Gene expression profiling of head and neck cancer

Warner, Giles C

GENE EXPRESSION PROFILING OF HEAD AND NECK CANCER

Doctor of Medicine

Giles C Warner MBBS MSc FRCS

2004

ABSTRACT:

The purpose of this study was to classify oral squamous cell carcinomas (OSCCs) based on their gene expression profiles, to identify differentially expressed genes in these cancers, and to correlate genetic deregulation with clinical-histopathological data and patient outcome. After conducting proof of principle experiments utilizing six head and neck squamous cell carcinomas (HNSCCs) cell lines, the gene expression profiles of 20 OSCCs and subsequently an additional 8 OSCCs were determined using cDNA microarrays containing 19,200 sequences and the Binary Tree-Structured Vector Quantization (BTSVQ) method of data analysis. Two sample clusters were identified in the group of 20 tumors that correlated with T3-T4 category of disease ($p=0.035$) and nodal metastasis ($p=0.035$). Sample clustering of 28 OSCCs and the 6 cell lines revealed a correlation with disease free survival. BTSVQ analysis identified a subset of 23 differentially expressed genes with the lowest quantization error scores in the cluster containing more advanced stage tumors from the 20 OSCC dataset. The expression of six of these differentially expressed genes was validated by quantitative real-time RT-PCR. Statistical analysis of quantitative real-time RT-PCR data was performed and, after Bonferroni correction, *CLDN1* ($p = 0.007$) over-expression was significantly correlated with the cluster containing more advanced stage tumors. Despite the clinical heterogeneity of OSCC, molecular subtyping by cDNA microarray analysis was able to identify distinct patterns of gene expression associated with relevant clinical parameters. The application of this methodology represents an advance in the classification of oral cavity tumors, and may ultimately aid in the development of more tailored therapies for oral carcinoma.

TABLE OF CONTENTS

CHAPTER 1: Introduction to Head and Neck Cancer

Cancers of the head and neck are the sixth most common cancers world wide.[1] There is an increasing incidence in developing countries and in patients with no risk factors.[2 3 4] The commonest histological type of tumor of the upper aerodigestive system is squamous cell carcinoma (HNSCC). Rapid developments in the molecular biology of these tumors have made it increasingly clear that understanding the molecular and cellular evolution of these tumors is crucial to the clinical and surgical oncologist to improve diagnostics, treatment regimens and survival in patients with this lethal disease.

1.1 Aetiology:

Cigarette smoking and alcohol consumption are the two strongest aetiological factors for the development of HNSCC both independently and synergistically.[5] Smoking unfiltered cigarettes carries a slightly higher risk than filtered cigarettes. The dose dependant relationship between smoking and incidence of HNSCC has been demonstrated in large autopsy and epidemiological studies, with the relative risk over non-smokers ranging from 2.4 for smokers of < 7 cigarettes per day to 16.4 for smokers of >25 per day.[6] However, cessation of smoking leads to a gradual reduction in risk by 70% after 10 years. Carcinogenic epoxides, which bind to DNA, are produced by the action of aryl-hydrocarbon hydroxylase on methylcholanthine and benzanthracene, substances released by the burning of tar.

The molecular genetic events caused by cigarette smoking have been elegantly demonstrated by two studies. *p*53 tumor suppressor gene mutations were detected in 42% (54/129) of consecutive HNSCC in one study.[7] However when the analysis accounted for social habits, *p*53 mutations were detected in the tumors of 58% of patients who smoked and consumed alcohol, 33 % of those who smoked but did not consume alcohol, and only 17% of those who neither smoked nor drank alcohol ($p = 0.001$). Another study

examining head and neck cancer cell lines showed that the p53 mutations were principally quanine and thymine transversions, a point mutation that can be affected by the interaction of DNA with benzopyrene.[8]

The association between alcohol and the development of upper aerodigestive tumors has proven more difficult to unravel. The most recent published study which analyzed this was a Danish population based study of 28 180 patients.[9] Over a follow – up period of 13.5 years, compared with non – drinkers, subjects who drank 7 – 21 units of beer or spirits but no wine had a relative risk of 3.0 of developing oropharyngeal or oesophageal cancer. However, subjects who drank a similar amount but who also drank wine as > 30% of their total alcohol intake had a relative risk of 0.5. The relative risk for subjects who consumed >21 units of alcohol excluding and including wine were 5.3 and 1.7, respectively. It seems from this study that wine drinkers may be at a lower risk of developing upper aerodigestive tumors than drinkers who have a similar intake of beer and spirits.

Given the strength of the causal relationship between chemical carcinogens and the development of HNSCC, a superfamily of enzymes known as the glutathione S-transferases (*GstP*), responsible for the detoxification of a wide range of xenobiotics, have been implemented as a pivotal role in HNSCC tumorgenesis.[10] Mice with a "knock out" deletion of the pi-class glutathione S-transferases had a four – fold increase in the development of skin papillomas after exposure to polycyclic aromatic hydrocarbon 7,12-dimethylbenzanthracene and the tumor-promoting agent 12-*O* –tetradecanoylphorbol-13-acetate.[11] In humans, increased expression of *GstP* has been reported in many tumors including oral cancers.[12] An important member of this family of enzymes, *GstP1*, is polymorphic in humans with allelic variants showing differences in their catalytic activities towards a range of carcinogens. It has been suggested that the carriage of one of the less active allelic variants might result in increased cancer susceptibility in an

individual. Indeed, it has been shown that individuals with oropharyngeal and laryngeal cancers were associated with significantly lower frequency of *GstP1 AA* polymorphism than controls, suggesting that the polymorphism at *GstP1* may mediate susceptibility to these cancers.[13]

The observation of synchronous and metachronous tumors in the upper aerodigestive system has given rise to the concept of "field cancerisation", a term coined to account for the presence of multicentric oral cancer. Various molecular studies have looked at whether this is the mechanism underlying the development of multiple primaries, or whether multiple primaries share the same clonal origin, but has subsequently migrated through the epithelium and later developed distinct genetic alterations such as in the case of multiple bladder tumors.[14] The detection of mutations and the overexpression of growth factors in histologically normal mucosa distant from the primary tumor tend to support the field cancerisation theory.[15]

Human Papilloma Virus (HPV) and Epstein – Barr virus (EBV) have principally been implicated in the pathogenesis of head and neck cancers. The E6 and E7 regions of high risk HPV types (6, 18 and 33) have been shown to inactivate the tumor suppressor gene products *p53* and *pRb,* potentiating neoplastic change and cell immortalisation.[16] Using polymerase chain reaction (PCR)-based analysis, HPV DNA has been detected in 15 – 62% of HNSCC.[17] Although the EBV has been detected in other head and neck cancers, EBV is especially associated with the development of nasopharyngeal carcinoma. Its genome is consistently found in nasopharyngeal tumor cells. Raised serum immunoglobulin A (Ig A) against the viral capsid antigen has been shown to precede clinical symptoms in nasopharyngeal cancer and the evidence suggests that patients in high risk regions with an elevated serum anti – EBV IgA level should have an endoscopic examination of the nasopharynx and blind biopsies to include the pharyngeal recesses.[18]

Site-specific carcinogens have been identified for some head and neck cancers. Nickel and chromate dust are principle inorganic chemicals which can cause lesions in the nose, larynx, lung and paranasal sinuses.[19] Clusters of cases in the hardwood industry led to the identification of hardwood dust as an aetiological factor in the development of adenocarcinomas of the paranasal sinuses.[20] Similar epidemiological evidence implicated nitrosamines in a salted fish diet below the age of ten in the development of nasopharyngeal carcinoma.[21]

The association of Patterson – Brown – Kelly syndrome (iron deficiency, glossitis, koilonychias and an upper oesophageal web) with postcricoid carcinoma may account for the relatively high incidence of both conditions in the UK and Scandinavia. This may also account for the fact that the postcricoid region is the only subsite of the hypopharynx where the incidence of carcinoma in women exceeds that of men and where the age of incidence is relatively young. The reported incidence of post – cricoid carcinoma in patients with oesophageal web is 4 – 16%. The decrease in prevalence of Patterson – Brown – Kelly syndrome in Sweden as a result of dietary changes and education has been mirrored by a similar fall in the incidence of post-cricoid carcinoma.

1.2 Family History:

Using multivariate analysis, a relative risk of 3.5 – 3.79 for HNSCC associated with family history has been demonstrated in a large case – controlled study from Brazil.[22] This risk rose to 7.89 in first-degree relatives of patients with multiple HNSCCs. Although the exact nature of the association is unclear, these findings do suggest that familial factors may be important in determining susceptibility to HNSCC. The high incidence of nasopharyngeal cancer in south – east China, coupled with its greatly increased incidence in first degree relatives of patients with nasopharyngeal cancer, has led to a linkage analysis pointing to a susceptibility gene near the human leucocyte antigen (HLA) cluster on chromosome 6p.[23] *In vitro* studies have demonstrated increased

mutagen sensitivity in HNSCC patients with a family history of HNSCC and further increases in those with two or more affected first-degree relatives or patients with multiple HNSCC.

1.3 Molecular Genetics of head and neck cancer:

An imbalance of the equilibrium between growth promoting and growth restraining signal transduction, and natural cell loss caused by proto-oncogene activation and tumor suppressor gene inactivation leads to an unbalanced mitogenic signal and consequent aberrant cell proliferation. This concept of genetic alteration leading to neoplastic phenotype and subsequent clonal expansion is the basis of the "clonal evolution model of tumor cell populations".

Recently, there has been an increasing wave of interest in the significance of apoptosis in the development of cancer and its treatment. The induction of apoptosis in untreated tumors is complex and may involve a number of factors including tumor necrosis factor – alpha and expression of the oncogenes c-*myc* and c-*fos* in certain circumstances. In malignant head and neck tumors, the mode of action of the proto – oncogene *bcl-2* has been shown to be a novel one by inhibiting apoptosis rather than stimulating cell proliferation.[24] Ionising radiation induces apoptosis in normal tissues and tumors to a variable extent and is related to an increase in the level of wild – type p53 protein in the cell. A number of co-ordinated genetic alterations affecting cell turnover is therefore required to initiate tumor formation and progression. The challenge has been the identification of these genetic events and their correlation to known events in the histological progression of tumors from normal mucosa, through dysplastic changes to invasive carcinoma.

1.3.1 Cytogenetic abnormalities, tumor suppressor genes and oncogenes:

Cytogenetic studies in solid tumors have been hampered by the difficulties in establishing short – term primary cultures and the erratically acquired chromosomal abnormalities in long-term cell lines that may have occurred in vitro, influenced by culture conditions. However, some studies have identified chromosomal areas in head and neck cancer consistently showing frequent breakpoints suggesting the location of putative tumor suppressor genes (including 3p21, 5q14, 8p11, 17p13 and 18q2) and gain or amplification implying the presence of putative proto – oncogenes at other sites (including 3q, 5p, 8q and 11q13). [25]

Chromosomal region 9p21 loss is the most common chromosomal aberration detected not only in head and neck cancer but in the majority of human cancers, occurring in over 70% of head and neck cancers.[26] Positional cloning strategies led to the identification and characterization of *p16* (MTS1 or CDKN2) as a candidate tumor suppressor gene in this area.[27] However, inactivation of this gene by point mutations is uncommon. Homozygous deletion and hypermethylation of the 5' promoter are the predominant modes of gene inactivation. Further support for the key role of *p16* as a tumor suppressor gene is given by the nature of its function as a potent cyclin/cyclin-dependant kinase (CDK) inhibitor involved in the G1-S cell checkpoint of the cell cycle aswell as the demonstration that transfection of full length cDNA of *p16* and *p16β* into head and neck cancer cell lines results in marked growth inhibition with cell cycle arrest in G1.[28] An important aspect of 9p21 deletions is that it is one of the earliest genetic events in the pathogenesis of all the cancers in which it has been implemented, specifically, non-small cell lung cancer, bladder cancer and head and neck cancer. The implication of this is that detection of 9p21 deletion or *p16* inactivation may have considerable potential in the definition of early and high-risk pre-malignant lesions.

About 60% of HNSCC show 3p deletion, with the specific regions being 3p14, 3p21.3, 3p22 and 3p25.[29] 3p14.2 is a site of chromosome fragility across which the FHIT gene

has been cloned. Different groups have reported mutations and aberrant transcripts of the FHIT gene in up to 65% of head and neck tumors.[30] However, the function of this gene, in particular to act as a tumor suppressor gene, remains unclear.

Loss of heterozygosity at chromosome 17p13 has been shown in over half of HNSCC and often correlates with *p53* inactivation.[31] While *p53* mutations are detected in early preinvasive lesions of the head and neck, their incidence increases with tumor progression. However, research appears to uncover ever more facets to a gene whose function is intimately related to that of other tumor suppressor genes, proto-oncogenes and growth factors. What is known, however, is that wild type p53 protein has a critical role in inducing G1 – arrest until repair has been affected; or if that is not possible, in directing the cell into an apoptotic state through transcriptional activation of such genes as *p21/WAF1*, *mdm2* and *bax*. Certain mutations of the p53 gene result in p53 protein that is transcriptionally inactive and therefore unable to execute this function, leading to propagation of acquired genetic alterations and outgrowth of malignant clones.

Allelic loss at 13q14 also occurs in over half of HNSCC.[32] The area of deletion includes the retinoblastoma gene (RB1) and is immediately adjacent to the hereditary breast cancer gene BRCA2. Immunohistochemical analysis of the RB1 protein product and mutation analysis of BRCA2 in head and neck tumors showing loss of heterozygosity at 13q14 have discounted both of these known tumor suppressor genes as putative tumor suppressor genes inactivated in this area and imply the presence of another tumor suppressor gene in this region.[33 34] The exclusion of BRCA2 is of particular interest as there is evidence that BRCA2 families show an excess of laryngeal cancer.

Consistent amplifications of 11q13 implicated PRAD1 (CCND1 or cyclin D1) as an important oncogene involved in HNSCC pathogenesis. Further work has shown that amplification of 11q13 is associated with increased expression of this gene. On a clinical level, PRAD1 overexpression correlates with tumor progression and is strongly

associated with reduced disease free survival in patients with operable head and neck cancer.[35 36]

The role of the *ras* family of oncogenes in HNSCC pathogenesis is unclear. While a high incidence of *ras* mutations has been reported in HNSCC in India, several studies have failed to support this in the Western world. This finding suggests that there may be a specific mutagenic effect of chewing tobacco or betel nuts upon the *ras* gene.[37 38]

1.3.2 Growth factors:

Polypeptide growth factors and their receptors mediate signals that stimulate cell division and growth in normal cells under physiological conditions. Over-expression of these growth factors and their receptors may promote pathologically excessive cell growth and, as such, they could be considered as protein products of proto – oncogenes, especially when oncogenic retroviruses carry genes (v-onc) showing strong homology to corresponding human genes (c-onc). An example of this is the homology between viral *v-erbB* and human epidermal growth factor receptor (EGF-R). Qualitative analysis of various growth factors in HNSCC at the DNA, mRNA and protein level has consistently demonstrated grossly elevated levels of epidermal growth factor (EGF), epidermal growth factor receptor (EGF-R) and transforming growth factor-α (TGF-α) in head and neck cancer as well as surrounding histologically normal mucosa.[39-41]

Over-expression of other growth factors and their receptors, such as platelet-derived growth factor (PDGF), fibroblast growth factors (FGF-1, FGF-2) and fibroblast growth factor receptor (FGF-R), has been demonstrated in head and neck cancers.[42] The growth factor-like molecule *HER-2/neu* has shown great promise in prognostic determination in breast cancer, while although these amplification have been shown in some HNSCC, it has failed to show any association with clinical parameters.[43 44] However, the growth factor receptors expressed on tumor cell surfaces are being extensively investigated with optimistic early results as tumor-specific targets for immunotherapy.

1.3.3 Telomeres and telomerase:

Telomerase activity has been demonstrated in 80% of oral cancers and around 50% or oral leukoplakia.[45] [46] While detection of telomerase activity in oral rinses from the head and neck cancer patients remain limited by technical problems of test sensitivity, the potential for antitelomerase drugs as a novel treatment remains promising.[47]

1.3.4 Molecular Progression:

Fearon and Vogelstein[48] proposed the pioneering genetic progression model for colorectal tumorgenesis in 1990. Specific genetic alterations were allied to each step of the well-established adenoma-carcinoma sequence in colorectal tumorgenesis. A similar model has been more difficult to establish in HNSCC but a number of models have been suggested by different groups from analyzing published allelotype data;[49] from microsatellite analysis of adjacent areas of histologically normal tissue, dysplasia, carcinoma *in situ* and invasive carcinoma;[50] [51] and from similar microsatellite analysis of tissue from individual patients biopsied at different times. Theses are largely in agreement with each other. Progression from normal squamous epithelium to hyperplastic epithelium is thought to be due to overexpression of EGF and EGFR.[52] Telomerase activation and *p16* inactivation causes progression to dysplasia. PRAD-1 amplification, *p53* inactivation and 3p deletion causes progression to carcinoma in-situ, and 4q, 5q, 8p and 13q deletion cause further progression to invasive carcinoma. Over-expression of matrix metalloproteinase is thought to be involved in metastasis.[53-56]

1.3.5 Tumor Immunology:

The successful development of a tumor depends on neoplastic cells escaping either recognition or destruction by the immune system. Such escape in HNSCC may arise from functional down-regulation of tumor infiltrating lymphocytes (TIL) by tumor-secreted factors such as prostaglandin E_2 and TGF-β.[57] The identification of tumor

specific antigens has both diagnostic and therapeutic implications. The monoclonal antibodies E48 and U36 bind to peptides expressed in HNSCC, stratified epithelium and transitional epithelium.[58-60] Technetium-99m-labelled E48 IgG has been used for clinical imaging of head and neck cancer[61] while both E48 and U36, coupled with iodine-131 and rhenium-186, respectively, have successfully been used for immunotargeting HNSCC in nude mice.[62] The clinical trial of immunotherapy showing the greatest promise in head and neck cancer has arisen from the observation that HNSCC cells secrete immunoregulatory cytokines which effect growth inhibition both directly and indirectly. Local treatment with IL-2 given by intratumoral injection in one trial of patients with recurrent HNSCC has been shown to give total or partial responses in around 30% of patients.[63] This may be related to an increase in T-cells and lymphokine activated killer cells within the tumors. Systemic treatment with cytokines such as IL-2, IL-12 and interferons has been limited by treatment-associated toxicity. Il-2 has been used to expand TIL *in vitro* followed by reintroduction of autologous IL-2 activated TIL.[64] Although such cellular adoptive immunotherapy showed early promise in animal models, it has yet to be translated into full clinical trials.

1.3.6 Angiogenesis:

Angiogenesis has been demonstrated immunohistochemically in a number of tumors, including HNSCC, with antibodies to vascular antigens such as factor VIII-related endothelial antigen.[65] Published data correlating the degree of tumoral microvessel density and clinical parameters in HNSCC is inconsistent, but there is good evidence that high microvessel density may predict a favorable response to radiotherapy, particularly in nasopharyngeal carcinomas.[66] Further promise for targeting angiogenesis comes from studies involving endostatin, an endogenous protein that is a potent inhibitor of endothelial cell proliferation, a process that underpins angiogenesis.[67] Convincing animal

data has shown that endostatin treatment of mice with implanted tumors resulted in complete resolution without recurrence after two to six weeks of treatment.[68]

1.4 Clinical applications

Perhaps the holy grail of the molecular oncologist is gene therapy. Even before the successful cloning of known tumor suppressor genes, a crude form of gene replacement therapy by introduction of whole chromosomes suspected of harboring putative tumor suppressor genes had been shown to suppress genes tumorigenicity *in vitro*. The introduction of wild type *p53* and *p16* into human head and neck cancer cells either by transfection *in vitro* or via a recombinant adenovirus *in vivo* in nude mice has consistently shown growth suppression in both circumstances.[69] Other genetic modulation treatments involve the use of virus-directed enzyme prodrug therapy (VDEPT). In head and neck cancer, this has been extensively investigated in experimental animal work. An example of this is the introduction of herpes simplex virus thymidine kinase gene into tumor cells, resulting in the expression of thymidine kinase in those cells and its conversion of gancyclovir into phosphorylated compound which halts DNA synthesis.[70] This type of treatment can be combined with IL-2 immunotherapy, which exhibits synergistic effects in animal models.

Recent advances in molecular biological research in cancer have improved our understanding of the genetic and cellular events underlying the development of cancers of the head and neck. Genetic determination of phenotype and prognosis in head and neck cancer will be more important and as more genes become identified and greater precision in such determination is achieved. Several cytogenetic breakpoints have been shown to be associated with radioresistance *in vitro*. Additionally, certain apoptotic markers show promise as indicators of resistance to chemotherapy.[71] As these become more defined, treatment strategies may be tailored to the tumor genotype. The identification of new genes and identification of the oncogenic process hold promise for

radical improvements to established methods of tumor evaluation and treatment as well as exciting new developments in novel therapeutic strategies.

CHAPTER 2: Introduction to microarrays

2.1 Introduction

The draft sequence of the human genome was published in 2001.[72] Researchers have now begun to unravel the functional relevance of these data. Many tools have been developed to meet this goal of which cDNA microarrays are the most promising for large-scale gene expression analysis. Traditional methods of gene investigation in cancer have relied on identifying single genetic alterations associated with disease. The strength of microarray analysis lies in the ability to analyze the expression of thousands of genes simultaneously. This allows a representation of the genetic activity of one target sample with or without comparison to a reference.[73] To date, microarray technology has altered the way scientists are approaching cancer research and may be important in optimizing treatment and improving outcome of patients with head and neck cancer. Although still in its infancy, it is likely to have direct application to diagnosis and prognosis in the near future.

2.2 Microarray Technology:

Not all genes are expressed at the same time or at the same level in the genome. Furthermore, the regulation of a given gene in relation to other genes ultimately results in normal cell function. Therefore, high throughput techniques were developed to detect and quantify thousands of genes at the same time. This resulted in the ability to perform simultaneous gene expression analysis, or "global gene profiling". Microarrays are platforms on which thousands of gene sequences are printed. Fluorescently labeled cDNA sequences from the sample of interest are hybridized to these targets. The fluorescent intensity of each hybridized sequence in the array is then read. The scanner that records the intensity value is linked to custom digital image analysis software, which produces a color-coded image of the array, and a quantitative value is recorded for each

target gene. Intensity of fluorescence correlates with expression of the gene for which the spot codes. A schematic of the basic microarray protocol is shown in Figure 1.

There are two ways in which microarray slides are manufactured. The main difference between these two techniques lies in the size and nature of the sequences used. In the first technique, a high-density array of 25mer oligonucleotides is synthesized *in situ* on a glass surface using combinatorial chemistry and photolithography.[74] [75] Affymetrix (Santa Clara, CA) manufactures this commercially available array, named "GeneChip®". In this technology, 10 pairs of DNA sequences represent each gene. The ability to identify single nucleotide polymorphisms and mutations is the major advantage to using this array[76], because it includes sequences presenting single base-pair differences or mutations. Lindblad-Toh *et al.*[77] performed a loss of heterozygosity (LOH) analysis of small-cell lung carcinomas using single polymorphism arrays. These arrays can be broadly applied to several fields of medicine and research such as medical diagnostics and pharmacogenomics. In addition, owing to the small sequence size, a greater density of probes is housed on the same array. Affymetrix has recently introduced arrays housing up to 40,000 sequences, as well as gene-specific arrays (e.g. "*TP53* GeneChip®"). Further details can be found at www.affymetrix.com. The standard Affymetrix protocol uses 5-10μg total RNA as starting material, from which biotinylated cRNA is synthesized. The use of less total RNA (2μg of starting total RNA) has been reported[78] with results equivalent to the 10μg required using cDNA arrays. Finally, these commercial arrays have high unit costs; however the production process ensures consistency between arrays with no technical variability.

The other technique uses cDNA sequences of known genes and Expressed Sequence Tags (ESTs). ESTs are sequences that are identified as potential mRNA, isolated, catalogued and cloned. The assignment of one EST to a particular gene is often done by sequence similarity. In the cDNA microarray technique, sequences are amplified by the

Polymerase Chain Reaction (PCR) and printed onto glass microscope slides using high-precision robotics. The size of amplified sequences is variable (500-2000 base pairs). There are currently several types of cDNA arrays available from different research institutions. cDNA arrays from the Microarray Centre of the University Health Network (www.microarrays.ca) were used in these experiments. The human cDNA clone sets printed on the slides are from the IMAGE Consortium (http://image.llnl.gov/).

Other methodologies are in current use or development. Hybridisation of mRNA to probe sequences has been carried out on a spherical surface using glass beads as a substrate rather than a flat surface. Okamoto et al describe a method that uses the same technology found in inkjet printers to eject a microdroplet of oligonucleotides onto a glass surface.[79] They fabricated a custom microarray using bubble jet technology and then used the microarrays to discriminate between 2 SCC cell lines. Other techniques, such as serial amplification of gene expression (SAGE), directly measure mRNA concentrations rather than the hybridization of mRNA to DNA probes.

2.3 Bioinformatics

Bioinformatics is defined as the application of computational techniques to biology, in particular to molecular biology. The goal is to provide computer-based methods for coping with and interpreting large volumes of diverse data obtained by high-throughput approaches. When applied to microarray experiments, it is a complex subject and represents a mathematical evaluation of vast amounts of gene expression data in order to decipher coherent gene and metabolic pathway function in a global sense. Computational methods, capable of accurately distinguishing between different biological or clinical categories, and identifying determining genes are needed to organize and interpret microarray data. Each quantified expression values must be first normalized before further computational programs are introduced. Subsequent to this various gene clustering algorithms are applied depending on the question to be answered.

2.3.1 Gene Expression Data Characteristics.

Gene expression data distributions can best be described by multivariate Gaussians.[80] Exploration of gene expression data sets is always problematic due to its inherent dispersion and missing values.[81] In order to apply a cluster analysis technique, it is important to explore the characteristics of gene expression data. The selection of a clustering technique should be based on how the technique relates to the characteristics of the data, and whether this technique is able to find any biologically significant clusters.[81] In experiments involving a test versus reference design, representing differentially expressed genes with a log ratio provides an equal spread between up- and down regulated genes, with a mean of zero for equally expressed genes. Although there are clear advantages to using log ratios to represent the normalized data, this introduces the problem of loss of information about the absolute gene expression level in the sample. This may represent a statistical shortcoming, since equal weights could be given to ratios that are based on widely different absolute levels of expression. The background intensity poses another set of issues to be addressed. First, there is no single established method for deciding what is the most appropriate area around the spot for measuring background. In addition, subtracting locally measured background from the spot intensity value can result in a nonsensical intensity measurement.

Preprocessing of the original data using an appropriate normalization technique can minimize various types of errors. Normalization should account for the dye bias, both in labeling and detection efficiencies, and for the nonlinear relationship between dye intensity and expression level.[81] It should also handle negative intensity values, which may result from background subtraction, as well as variance within and among experiments. In the worst case, normalization may introduce more error into the data than it removes (as is often the case with background subtraction), thus further decreasing the quality and reliability of the results. Normalize Suite software package

(www.utoronto.ca/cancyto) normalizes Cy5 and Cy3 fluorescence intensities with or without background subtraction, across the entire array as well as across individual sub grids. Additionally, preceding normalization, filters are applied to remove spots below user-specified thresholds for fluorescence intensity, foreground-to-background ratio, and/or spot diameter size. After normalization and filtering, duplicate spots are averaged together and saved to disk. Two or more replicates of an experiment can be grouped together as a project. Thus, upon normalization of replicate arrays of the same experiment, the saved normalized files are averaged together using the 'project' software. Experiments where the fluorophore-labeling has been reversed are easily integrated with existing data. Finally, the 'clusterp' software identifies all the project files and saves these in a format that can be used by the Eisen Cluster software package [82] (http://rana.lbl.gov) or other clustering systems..

There are many additional issues related to microarray data generation, with numerous potential sources of error from the biological and technical side. In addition, the concentration of cDNA used in hybridization is non-linearly related to the signal intensity [81], which complicates any experimental conclusions. However, these issues are beyond the scope of this chapter.

2.3.2 Gene Clustering

Current analysis of microarray data involves applying both statistical and machine learning techniques, such as hierarchical clustering[82], self-organizing maps [83]or *K*-means clustering[84] to organize genes and patient samples into meaningful groups. These methods have been extensively used in most microarray studies.[85-88] Hierarchical clustering is an agglomerative approach widely applied to databases of gene expression profiles from patient samples. It can delineate clinically relevant patterns in diverse cancers. This method successively merges individual data points into a tree structure called a dendrogram, based on their similarity.

Self-organizing maps are well suited for exploratory data analysis and are considered superior to hierarchical clustering when analyzing data containing outliers and sources of systemic error. Partial structure is imposed on the data, and then adjusted according to the results. The input is the raw expression data, and the output is a series of maps representing gene expression.

Other approaches, such as Bayesian clustering, require availability of prior distributions of the data, while *K*-means clustering requires that the researcher determine *K*, which specifies the number of clusters in the data. It is becoming clear that the high complexity of expression data poses a challenge to existing tools. It has been shown that a combination of diverse basic approaches is a suitable manner to analyze of microarray data.[89] Most existing tools have been developed for relational types of data, which typically have a large number of instances but low complexity. Thus, high complexity causes many existing tools to fail or provides outcomes with limited usefulness. New tools must be flexible enough to support the diverse tasks associated with clinically relevant genomic research

2.3.3 Binary Tree Structured Vector Quantisation

One novel approach to microarray data analysis is the Binary Tree-Structured Vector Quantization (BTSVQ) system [90]. The algorithm uses self-organizing maps (SOM) [83] and partitive k -means clustering [84] in a complementary fashion. It applies the vector quantization and self-organization capabilities of SOMs to find significant gene centers in gene space, which is characterized by high dimensionality and a large number of clusters. BTSVQ uses the effectiveness of k -means in sample space, which is characterized by medium dimensionality and a low number of clusters. The SOM algorithm finds gene centers by vector quantization and places genes with similar expression in neighboring units in the two dimensional map grid. Thus, SOM preserves similarity relationship in gene space. This quantization and re-arrangement of genes

ensures that quantized genes would have similar expression for similar samples. Thus, the component planes of similar samples would look similar. BTSVQ superimposes results from the two complementary clustering approaches – the highest confidence in the clustering result is achieved when samples with visually similar component planes are placed in the same child of the cluster tree.

Beacause of the inherent dispersion and skewness of microarray data, and its closeness to multivariate Gaussians[80], BTSVQ adopts an iterative partitive clustering approach which extends the TSVQ approach. [91] The algorithm partitions data using the standard *k* - means algorithm in sample space, where *k* is kept constant at 2. Iteratively applying the algorithm and using evaluation of variance as a stopping criterion it generates a binary tree. The SOM algorithm is then used to cluster the gene space. The cluster structure in gene space is visualized using component planes of the already computed SOM. An outline of the binary tree-structured vector quantization algorithm (BTSVQ) is given below.

1. Binary Tree generation:

   (a) Start with the whole data set in a single cluster (parent).

   (b) Partition the data set (experiment space) into two subsets using standard *k*-means (simple Euclidean distance measure is used)

   (c) Compute variance of both subsets (children).

   - If variance of parent from step 1a is less than the variance of child from step 1b, stop further partitioning of that child.
   - Repeat step 1b for all remaining children, taking each child as a parent.

2. Visualization:

(a) Generate component planes of SOM for all samples.

(b) Arrange the component planes at nodes of the B-tree generated in step 1.

Component planes of SOMs are the planes of Voronoi Tessellations[92], each representing a sample in a microarray experiment (i.e., gene expression profile of a tumor). Figure 2 represents quantized gene expression visualizations by component planes of a SOM. A hexagonal map unit in a component plane represents a gene selected by the SOM algorithm, which is a center of a cluster in the data. The color of the map unit represents the expression value of the gene associated with that unit. Thus, similar gene expression profiles correspond to similar color patterns of the component plane.

BTSVQ has been used to identify clinically relevant molecular sub-classifications of tumors from gene expression data. One analysis revealed a number of novel observations on the pathobiology of lung cancer, most importantly, the presence of molecular subtypes of non-small cell lung cancer correlating with differences in rates of tumor recurrence and disease-free survival not evident from agglomerative clustering of the same data.[89] The BTSVQ algorithm delineated a subgroup of patients using all available gene expression data. In this patient cluster, a significant proportion of selected patients were characterized by an early recurrence of disease, indicative of more aggressive tumor biology. 14/17 of these patients had evidence of early disease recurrence on clinical follow-up, compared to 2/7 in an analogous subgroup on the opposite arm of the cluster tree.

This type of system is ideally suited to analyze data from genetically heterogeneous samples such as those obtained from Head and Neck Cancer. This data is often skewed with many outliers and variable expression values requiring greater computational input

to identify significant gene expression patterns. Genetic profiles for each sample are readily visualized in the SOM clustered at the nodes of the binary tree thus creating a simple and biologically intuitive way for interpreting microarray data. Additionally, this system has been shown to be less sensitive to preprocessing and normalization, and successfully revealed clinically relevant clusters in three large publicly available lung microarray data sets.[90] Tibshirani et al found that this type of clustering method successfully organized genes and experiments and generated useful and visible data organization.[93]

BTSVQ selects a unique set of genes that best discriminates each sample cluster positioned at each node of the binary tree. Selection is based on the gene quantization error (QE). This is the probability of the gene having a similar expression value across all samples in that cluster thus providing a measure of the ability of a gene to define clusters of samples. The lower the QE, the higher the probability. In conventional data analysis such as hierarchical clustering, genes that fall below a certain fold change of expression are excluded from the analysis, thus selecting only highly differentially expressed genes. Based on a Cox proportional hazards model, Wigle et al showed that many genes that were below a certain fold change were capable of class distinction between groups of lung cancer patients with early recurrence versus no recurrence.[94] Thus gene selection by QE aims to avoid the biases introduced by arbitrary threshold values.

2.4 Applications

Microarrays can be applied to both research and clinical settings. Below are some areas where microarray analysis has already made an impact.

Expression profiling:

Structure and function of cells and tissues is determined by the selective, differential and collective expression of many genes. Measurement of global RNA expression represents

a tissue state of activity and differentiation. This allows the direct comparison of genetic activity of cancer cells with their normal counterparts in any tissue. Global expression profiles allow us to build gene maps showing how genes are functionally related to each other within the genome. Ross *et al.*[95] have recently described the pattern of expression of 8000 genes in each of the 60 tumor cell lines used in the NCI screen for anticancer drugs. Each tumor type and its tissue of origin expressed distinctive gene subsets ("molecular signatures"), which correlated with known function. Thus, for example, melanoma cell lines prominently expressed genes associated with melanin metabolism. This may help classify the lineages of complex clinical samples such as poorly differentiated tumors and metastases of unknown primary origin.

Molecular Tumor Classification:

Improvements in tumor classification are central to the development of novel and individualized therapeutic approaches. Histologically indistinguishable tumors often show differences in clinical behaviour, and subclassification of these tumors based on their molecular profiles may help to explain why these tumors respond differently to treatment. In a landmark study, Golub et al. [96] applied microarray technology to develop innovative classifications to leukaemias, using microarray analysis based on "neighborhood analysis" and the utilization of tumor class predictors. This strategy was able to distinguish between acute myeloid leukeamia and acute lymphocytic leukeamia without supervisory analysis. Other groups have also used gene expression analysis to classify, at the molecular level, breast tumors,[97-99] B-cell lymphoma,[100] cutaneous melanoma,[101] alveolar rhabdomyosarcoma[102] and lung adenocarcinoma.[103] [104] Likewise, in a recent study analyzing molecular profiles of 50 nonneoplastic and neoplastic prostate samples, Dhanasekaran et al established signature expression profiles of healthy prostate, benign prostatic hyperplasia, localized prostate cancer, and metastatic prostate cancer.[105] These studies established the feasibility of combining large scale molecular analysis of

expression profiles with classic morphologic and clinical methods of staging and grading cancer for better diagnosis and outcome prediction.

Mutation detection:

Human diseases are frequently associated with mutations that affect specific genes. For example, over 400 mutations have been found in the *BRCA1* gene.[106] These mutations can be detected by creating 25mer oligonucleotide arrays with large numbers of single nucleotide variations. This allows rapid and precise identification of mutations or variation in individuals, families or populations. This process could be scaled up to simultaneously identify mutations or variants in large fragments of the human genome.

Pharmacogenomics:

Microarrays offer a number of new approaches to drug development and the prediction of therapeutic efficacy. Many genes in a cell can affect its response to therapeutic agents, in relation to drug exclusion, resistance, metabolism and DNA repair. The study of gene expression and genetic variation may allow the characterization of each cell, tissue and tumor in terms of predicted drug sensitivity, drug toxicity and development of resistance to therapeutic agents.[107] This analysis of individual variation is the basis of the new science of pharmacogenomics.

Gene copy number analysis:

Many tumors contain abnormal copy number of genes through processes such as chromosomal and gene deletion, amplification and aneuploidy. Quantifying the number of gene copies per cell can be helpful in understanding a disease process. Pollack *et al* were the first to describe microarray-comparative genomic hybridization (CGH) based techniques using cDNA targets.[108] This methodology consists of hybridizing target sample genomic DNA to cDNA arrays to identify copy number imbalances.[109] Recently, high-resolution profiling using array-CGH was used to study gene copy number changes in human tumors such as schwannoma[110] and neuroblastoma,[111] and for detection of

occult micrometastatic tumor cells.[112] Microarray-CGH thus allows the study of copy number changes across the entire genome of any one sample.

2.5 Microarrays in head and neck cancer

Head and neck cancer is a disease of considerable mortality. The overall prognosis of patients has not changed significantly in the past decade, with the 5-year survival rate remaining static at approximately 50%. It is hoped that novel molecular advances such as microarray technology may help in the understanding of this disease and ultimately lead to improvements in diagnosis, treatment strategies and outcome.

Microarray technology has recently begun to be used extensively in head and neck cancer research. Most studies relate to identification of differentially expressed genes between normal and malignant tissue, or between known subtypes of the disease. Belbin *et al.*[113] used hierarchical clustering to identify a subset of differentially expressed genes to classify 17 head and neck tumors. Overall, the molecular classification of these tumors was a better predictor of disease-free survival than clinical and pathological parameters. Villaret *et al.*[114] used a cDNA array consisting of 985 cDNA probes to analyze gene expression differences between 22 normal tissue and 16 frozen section samples of HNSCC from multiple sites. 13 genes were differentially overexpressed in the tumor tissue, four of which were previously unidentified. A larger set of genes was analyzed using five paired cases of normal vs. HNSCC.[115] They demonstrated differential expression of various oncogenes, tumor suppressor genes and other genes, many involving cell cycle regulation and cell signaling. Several of these have not been previously reported as differentially expressed in head and neck cancer. Leethanakul et al used a cDNA array housing 588 known human cancer-related genes and nine housekeeping genes to examine five LCM-derived normal and head and neck cancer samples.[116] Down regulation of cytokeratins and an increased expression of several signal transduction and cell cycle-regulatory genes were noted. Recently, Squire et al have

identified recurrent chromosomal alterations by CGH and SKY and correlated these with expression array analysis in head and neck tumor cell lines and primary tongue tumors.[117] These and other studies clearly demonstrate the potential power and use of this technology.

Despite its potential, microarray technology is still relatively new and expensive, thus limiting the type and extent of research possible. Initial experiments have been criticized as having no underlying hypothesis. Additionally, the choice of array may lead to the identification of many genes irrelevant to the disease. Furthermore, validation of genetic deregulation by alternative techniques, such as immunohistochemistry, tissue arrays, fluorescent *in situ* hybridization (FISH), quantitative real time PCR, Western or Northern Blot are crucial in adding credibility and confidence to the data generated using microarrays.

Tissue arrays involve taking six cores of 5 μm tissue from at least one hundred paraffin embedded tumor samples, and using these to create a new paraffin block. Sections from this new paraffin block can then be cut and interrogated with antibodies to identify which tumors express the protein of interest. Issues relating to accurate tissue sampling, standardization of staining patterns and number of sections required from each tumor are still problematic, but this technique allows the detection of the protein product of genes of interest simultaneously in a large number of tumor samples and is thus very powerful.

Quantitative real-time PCR relies on the quantitation of the expression level of one gene in tumor samples in comparison to a reference sample. A reference gene (e.g. ribosomal RNA, β-*actin*, *GAPDH*), which should not vary in transcript levels in different tissues or states of differentiation, is also used as an amplification control. PCR can be performed using fluorescent dyes or probes and allows a quantitative detection of specific product accumulation during each PCR cycle. Relative expression levels or gene copy numbers are directly related to the amount of starting target cDNA or DNA and are determined by

measuring the cycle number (Ct) at which fluorescence intensity becomes statistically significant.

2.6 Consolidating microarray data

Establishing a consensus on the best overall method of microarray analysis is likely to prove challenging. Variability in RNA extraction, choice of array, hybridization and labeling protocols, and analysis software leads to difficulty in comparison between studies conducted in different laboratories. When one combines this with the genetic heterogeneity of most solid tumors and the vast amount of information generated by microarrays, it is a concern that, instead of refining our current knowledge of functional genomics, we will in fact be unable to condense the information into a usable form. To help address this problem in head and neck cancer, the Head and Neck Cancer Genome Anatomy Project (www.cgap.nci.nih) has been established as a joint program between the National Cancer Institute (NCI) and the National Institute of Dental and Craniofacial Research (NIDCR). The aim of this Project is to create libraries of genes that are consistently expressed in tumors of the head and neck, which will therefore serve as a global reference for any investigator. This interactive website allows one to identify genes from the accession number (Unigene ID) and generates a large amount of information, including gene function, interacting pathways, and chromosomal localization. Genetic databases such as these will allow collation of all the raw data for effective evaluation and interpretation, thus assisting in multi institutional genomic and proteomic initiatives and collaborations.

CHAPTER 3: Gene expression profiling of HNSCC cell lines and Oral Squamous Cell Carinoma tumor samples

3.1 Introduction:

Oral squamous cell carcinoma (OSCC) is a subset of head and neck cancer. It is the 10th most common malignancy, comprising 2-3% of all new malignancies diagnosed in the United States.[118] Etiological factors include consumption of tobacco and alcohol, poor oral hygiene, nutritional deficiencies and possibly certain viruses.[119-121] Despite progress in treatment over the last few decades, the average five-year survival rate of 50% has remained unchanged resulting in approximately 13,000 deaths yearly in the United States.[122] This poor prognosis is associated with several factors including the high incidence of regional metastasis at presentation.[123] [124] Additionally, the lack of suitable markers for early detection, late presentation, insensitivity to available treatment and our limited understanding of the molecular mechanisms responsible for this disease all contribute to poor outcome.

Based on the hypothesis that metastatic potential is determined mainly by the genetic properties of the primary tumor, a number of genes have been studied in primary carcinomas in an attempt to determine biological markers that may aid in classification and prognosis. However, the molecular heterogeneity seen within individual diagnostic categories has meant that no potential marker identified thus far has proven to be a sufficiently strong indicator of treatment response or prognosis to be used in the clinical setting.[125] [126]

Disease staging is largely based on anatomical considerations gained from clinical examination and imaging techniques. Differences in treatment response and outcome of patients with stage-matched tumors suggest that the current staging systems are limited in their ability to predict loco-regional control and survival for head and neck cancer

patients. For example, even when the neck has been clinically staged as N_0 (*i.e.* no evidence of regional neck nodal disease) the incidence of micrometastases in postoperative histopathological sections of neck dissections is approximately 30%.[127 128] Accurate staging is therefore crucial to ensure that optimal therapeutic strategies are implemented in order to improve patient prognosis and survival.

Recently, global gene expression profiles generated by cDNA microarray analysis have been used as a powerful tool to classify different types of human cancers.[94 96 129] It has become possible to predict the subtype of the disease without previous biological knowledge of the tumor by using genes that best discriminate between different forms of the disease. Thus, correlations between the genetic profiles of HNSCC cell lines and primary OSCC samples with clinical parameters, including nodal status, may provide the basis for a molecular classification system.[130]

Several studies have used cDNA microarray analysis to classify HNSCCs according to their gene expression profiles, and to identify genes involved in tumor development and/or progression.[113-117 131 132] Despite all of these studies, there are still few molecular markers that are clearly associated with prognosis and disease mechanisms. In addition, many studies focus on individual genes and have not analyzed clusters of genes that are similarly under- or over-expressed in groups of samples. Such clusters of genes may be more clearly associated with disease mechanisms and/or clinical parameters.

In this study, cDNA microarrays were used to classify 6 HNSCC cell lines, 20 OSCCs and subsequently a combination of the 6 HNSCCs and 28 primary OSCCs according to their gene expression profiles. Most microarray studies have been performed using tumors from different sites in the head and neck. However, by analyzing a subset of head and neck tumors, the aim was to minimize the genetic differences that may be present in anatomically diverse tumors. To the author's knowledge, this was the first report of cDNA microarray studies that have used as many samples from the same anatomical site

in the head and neck. The results suggested that gene expression profiling could determine deregulation in multiple genes potentially associated with advanced stage disease. These findings may help to form the basis for a molecular classification of HNSCC, thus improving diagnosis, treatment and outcome for patients with this carcinoma.

3.2 Material and Methods:

3.2.1 Cell Lines

The cell lines UTSCC-9, UTSCC-24A, UTSCC-24B, UTSCC-34, UTSCC-60A and UTSCC-60B were provided by Dr. Reidar Grénman, Dept. of Otolaryngology, Turku University Central Hospital, Turku, Finland. Clinical data of patients from whom the tumor cell lines were derived are shown in Table I. Two cell lines were derived from primary tongue carcinomas (UTSCC-9 and 24A), one was from a tonsillar carcinoma (UTSCC-60A) and one was from a supraglottic laryngeal cancer (UTSCC-34). Cell lines UTSCC-24B and UTSCC-60B were from neck nodal metastases of the primary tumors from which cell lines UTSCC-24A and UTSCC-60A were derived.

3.2.2 RNA isolation from cell lines

Cell lines were maintained in keratinocyte growth medium. Following aspiration of the media, cells were washed once with Phosphate Buffered Saline (PBS). 5mls of PBS was added and the cells scraped from the plate. The cell suspension was transferred to a 50 ml polypropylene conical-bottom tube (Falcon Cat# 352070). The plate was washed with an additional 1ml PBS and the suspension added to the tube. A pellet of cells was produced by centrifugation at 2,300 rpm for 3 minutes at 4°C and the supernatant discarded. 2ml TRIZOL per ~$2x10^7$ cells (approximately one 150 mm plate of keratinocytes) was added to the pellet and the suspension passed through an 18 gauge syringe several times to disrupt the pellet. The sample was incubated at room temperature for 5 minutes. 0.4ml of chloroform (0.2ml/1ml TRIZOL) was added and shaken vigorously for 1 minute then incubated at room temperature for 2 minutes 30 seconds.

Cellular debris was removed by centrifugation at 4,000 rpm for 15min at 4°C. The supernatant was transferred to 1.2 ml microfuge tubes (0.5ml/tube) and an equal volume

of isopropranol added to precipitate the RNA. The sample was incubated at room temperature for 15 minutes then centrifuged at 15,000 rpm for 15 minutes to pellet the RNA. The supernatant was discarded and the pellet re-suspended in 70% ethanol. The RNA was stored in 70% ethanol at -20°C until use. Prior to use, the sample was centrifuged at 15,000 rpm for 15 minutes at 4°C and the supernatant discarded. The pellet was re-suspended in diethylpyrocarbonate (DEPC) treated water for labeling. The quantity and purity of the RNA was evaluated as per the tumor samples.

3.2.3 Tumor Samples:

The initial analysis involved twenty oral cavity tumor samples. Subsequent to this a further 8 samples were obtained, increasing the study number to 28. All samples were obtained at the time of surgery from the Toronto General Hospital, Toronto, Ontario, Canada. Tissues were snap frozen in liquid nitrogen within 30 minutes of resection and stored at -80°C in a tumor bank. To avoid contamination from normal stromal tissue, corresponding H&E-stained sections were examined by a board certified histopathologist to confirm the presence of at least 80% tumor cells in representative sections. Many studies are now using laser capture micro-dissection (LCM).[115] This technology increases sample purity and thus has the potential of generating more accurate gene expression data without stromal contamination. However, this technique harvests only a small amount of tissue and consequently only small amounts of RNA are obtained. RNA amplification techniques optimized for LCM samples have been commercially developed. Linear RNA amplification protocols have been used to amplify as little as 10ng total cellular RNA, corresponding to the amount that can be obtained from 10^3-10^4 cells.[133] [134] However, these techniques often do not result in representative amplification of all transcripts present in the original sample. Iscove *et al.*[135] described a rapid reverse-transcriptase PCR procedure that preserves RNA abundance relationships after amplification.[13] Other PCR-based amplification methods have been reported to retain the expression information for

most genes in microarray experiments.[136] [137] Because of the limitations in laser capture microdissection, the possibility of inaccurate RNA amplification and the added time required to process the material in this way, the method of insuring purity of samples by examination of corresponding histological slides was thought to be satisfactory. Other studies have used this technique to procure pure specimens resulting in informative microarray data.[94]

3.2.4 RNA isolation from tumor samples:

Tissues were homogenized and total RNA isolated by a single extraction method using an acid guanidinium thiocyanate-phenol-chloroform mixture. The method provides a pure preparation of undegraded RNA in high yield and can be completed within 4 h. It is particularly useful for processing large numbers of samples and for isolation of RNA from minute quantities of cells or tissue samples.[138] Following homogenization, 0.1ml of 2M sodium acetate (pH4), 1ml of phenol and 0.2ml of chloroform-isoamyl alcohol mixture (49:1) were sequentially added, with thorough mixing by inversion after addition of each reagent. The final suspension was shaken vigorously for 10 seconds and cooled on ice for 15 minutes. The samples were centrifuged at 10,000g for 20 minutes at 4°C. The aqueous phase was transferred to a fresh tube, mixed with 1ml of isopropranol, and then placed at -20° for at least 1 hour to precipitate RNA. Sedimentation at 10,000g for 20 minutes was again performed and the resulting RNA pellet was dissolved in 0.3mls of 0.2M sodium acetate, phenol and chloroform, transferred into a 1.5ml Eppendorf tube, and precipitated with 1vol of isopropanol at -20° for 1 hour. After centrifugation for 10 minute at 4°C, the RNA pellet was resuspended in 75% ethanol, sedimented, vacuum dried (15 minute), and dissolved in 50µl of 0.5% SDS at 65° for 10 minutes. RNA was evaluated by spectrophotometry and by electrophoresis on a 1% denaturing formaldehyde agarose gel by measuring A260 and A280. RNA was prepared to a final

concentration of 1μg/μL. RNA from all 28 tumor samples was found to be suitable for cDNA microarray analysis.

3.2.5 Patient Information:

Patient clinical data and tumor characteristics are described in Table II. Since consumption of tobacco and alcohol are important etiological factors in the development of head and neck cancer, only patients with a history of smoking and alcohol consumption were included in the study. Additionally, patients with a prior history of treatment HNSCC (surgery, radiotherapy, chemotherapy) were excluded. Although human papilloma infection plays an etiological role in HNSCC development in 10 – 20% of cases, human papillomavirus analysis was not performed, as it was not the primary focus of the study.[139] This work was approved by the local Ethics Committee and informed consent was obtained from all the patients prior to sample collection. The tumor bank database was examined to obtain clinical and histopathological information for each patient, including age, sex, history of consumption of tobacco and alcohol, disease stage, recurrence and outcome. This was often incomplete, and so in addition, patients' notes were accessed from medical records and audited to ensure accuracy of data. Some follow up information was obtained from the Ontario Cancer Registry. Since the Toronto General Hospital is a tertiary referral center for the whole of Ontario, patients were often referred from remote areas of the state. They received treatment within the specialist head and neck unit and were subsequently followed up locally. Therefore, in selected cases, full outcome details were obtained from direct communication with local hospitals or patients' general practitioner. Tumors were staged according to the current TNM classification as recommended by the International Union Against Cancer (UICC, 2002). [131]

3.2.6 cDNA Microarrays

Human 19K2 and 19K3 cDNA microarrays were purchased from the Microarray Centre, Ontario Cancer Institute, University Health Network, Toronto, Ontario, Canada. The 19K clone set information is available at the website microarrays.ca/support/download.html-uman19k3.zip. These arrays house 19,200 features including 18,980 human cDNAs and 220 positive and negative control features. The sequences on the 19K microarray are arranged on two glass slides (parts A and B), each one containing 32 sub grids with 24x25 features spotted over an approximate area of 18 x 36mm^2. All features are placed in duplicate, for a total of 38,400 spots. Two series of microarray experiments, one using cell lines derived from HNSCC and the other using primary OSCC tumor samples were performed as described below. Sequence verification studies performed by the Clinical Genomics Centre, Toronto, Ontario, Canada, allowed identification by BLAST of a small percentage of cDNAs contaiminated with mitochondrial DNA. Identification of such sequences allowed appropriate filtering of features during data analysis.

3.2.7 Labeling of cDNA and Hybridization to Arrays

cDNA microarray labelling protocols involve either direct or indirect labeling of cDNA with fluorescent dyes. Direct labelling is a fast, one - step method, while indirect labelling is a two step procedure involving AAdUTP in the reverse transcription reaction to incorporate an amine-modified nucleotide and subsequent coupling with monofunctional forms of Cy3 and Cy5 dyes which will react with the amine-modified cDNA. Indirect labeling uses less RNA, takes three days to process, and may give higher intensity spots. The direct labeling protocol (available from the Microarray Center of the University Health Network, Toronto, Ontario, Canada) was used for these experiments. The direct labeling protocol was chosen since it is a simpler, faster technique and has

been shown to work consistently by the University Health Network (UHN) Microarray Centre, Toronto, the institution from where the microarray slides were purchased.

Reverse Transciption:

Total RNA (10μg) from each tumor sample or cell line was used to synthesize cDNA with 400 units of the Superscript II reverse transcriptase enzyme (Invitrogen/Life Technologies, Inc., Burlington, CA). Reactions were performed in a buffer containing 5X first strand buffer, 100μM AncT primer (5'-TTTTTTTTTTTTTTTTTTTTTTVN-3'), 0.1M DTT, 20mM dAGT, 2mM dCTP and Cy3-dCTP/Cy5-dCTP fluorescent dyes. For 10-20 μg of Qiagen purified total RNA, the reverse transcription reaction volume was 40 μl. The following volumes were combined on ice.

8.0 μL 5X First Strand reaction buffer (Superscript II, Invitrogen)

1.5 μL AncT mRNA primer (5'-T20VN, 100 pmol/_l)

3.0 μL dNTP (-dCTP) (6.67 mM each of dATP, dGTP, dTTP)

1.0 μL 2 mM dCTP

1.0 μL 1 mM Cyanine 3 or Cyanine 5-dCTP (NEN)

4.0 μL 0.1 M DTT

0.1 – 20 μg RNA (0.1-0.5 μg mRNA or 10-20 μg total RNA)

0.5 – 1.0 ng Control RNA

40 μL dH20 (Sigma)

All samples analyzed were referenced against Human Universal RNA (Stratagene, Vancouver, Canada). This RNA is composed of equal quantities of total RNA from a pool of 10 tumor cell lines derived from the mammary gland, liver, cervix, testis, brain, skin, B and T – cells, expressing a wide range of genes. This RNA provided a denominator for accurate ratio discrimination for microarray analysis. Other studies use cell lines derived from normal mucosa [117] or matched normal tissue from the same

patient[140] as a reference. Each experiment was performed in duplicate with reciprocal fluorochrome labeling.

The tumor and reference samples were labeled with either Cyanine 3 or Cyanine 5. Reactions were kept in the dark, as the dyes are light sensitive. The incubator was covered with aluminum foil to create a cover preventing light from photobleaching the fluorochromes. The reaction mixture was heated to 65°C for 5 minutes and then placed at 42°C for 5 minutes.

A semi-quantitative method was used to estimate incorporation. The reactions were spun to pull down condensate from the lid of the tubes. 1 µL of each Cyanine 5 labeling reaction was isolated. The 1 µL aliquot was run on a standard 1% agarose gel, using only 5% glycerol as loading buffer, for 30 minutes at 100 V/cm of gel. The gel was exposed to a STORM phosphoimager using the red screen. Using ImageQuant, the amount of unincorporated Cyanine5-dCTP vs the DNA smear (densitometry) was measured. A yield of 3-4% or higher incorporation was considered satisfactory.

Stop Reaction:

RNA was hydrolyzed by adding 4µl of 50 mM EDTA and 2µl of 10N NaOH, and incubated at 65°C for 20 min. Samples were then neutralized with 4µl of 5M acetic acid. A small aliquot of the reaction was pipetted onto pH paper to ensure the solution was neutralized.

Purification:

Reactions were purified via centrifugation through a 30-µm micropore column (Amicon® Microcon® PCR Centrifugal Filter Devices, Millipore, Bedford, MA), to remove unbound dye. The Microcon®-PCR sample reservoir was inserted into one of the two vials provided. 400 µL of nuclease-free water was pipetted into the sample reservoir without touching the membrane with the pipette tip. 100 µL sample was added to the reservoir and the cap closed. Occasionally, the volumes had to be adjusted to get a final

volume of 500 µL. The assembly was placed in a centrifuge, the cap strap aligned toward the center of the rotor, and spun at 1000xg for 15 minutes. The column was inspected to see that most of the water has gone through the membrane. If water was sitting on the membrane, the column was spun again for one minute. When the water had gone through, the assembly was removed to a clean tube and 5 µL of nuclease-free water carefully added to the reservoir (avoiding touching the membrane surface) and left for 30 seconds at room temperature. The reservoir was inverted in the tube and spun at 1000g for 2 minutes. Approximately 5-6 µL of sample was collected. The Cy3 and Cy5 reactions were combined for each set of reference and tumor samples.

Prehybridisation:

Arrays were prehybridized for 1 hour at 37°C in 50µl of DIGeasy hybridization buffer (Roche, Laval, Qc, Canada) mixed with 5µl of 10mg/ml yeast tRNA, 5µl of 10mg/ml salmon sperm DNA, and 5µl of 10% BSA in humid hybridization chambers. Slides were then washed in double de-ionized water and spin dried prior to hybridization with the labeled probes. This step resulted in significantly less background fluorescence.

Hybridisation:

Probes were suspended in 80µl of DIGeasy hybridization buffer (Roche, Laval, Qc, Canada) combined with 5µl of 10 mg/ml yeast tRNA, 5µl of 10mg/ml salmon sperm DNA, denatured for 2 min at 65°C, and then cooled to room temperature. Since the microarray slides were designed in pairs (Part A and Part B each containing 9,500 sequences in duplicate), a "face-to-face" hybridization was used. The two slides that make up the pair were faced together, slightly offset to create a lip along one edge. The barcodes on the arrays allowed a small space between the slides. The hybridization solution was carefully applied along the lip, allowing the solution to evenly occupy the space between the slides. This technique allowed a "bubble-free" hybridization. The slides were carefully placed into hybridization chambers. These were plastic microscope

slide boxes containing a small amount of DIG Easy Hyb solution in the bottom to keep a humid environment. Clean plain microscope slides were placed at every second or third slide position in the slide box to create rails. Each hybridization chamber held two hybridization slides. The lid was carefully placed onto the box and the box then wrapped with plastic wrap. The boxes were incubated on a level surface in a 37 °C incubator overnight (about 18 hours).

Washing:

The hybridization chambers were disassembled and the pairs of slides carefully worked free of one another by immersing in 1X SSC and sliding the arrays past one another. When all of the arrays had been removed from the hybridization chambers, they were washed 3 times for 15 minutes each at 50°C in clean slide staining boxes containing pre-warmed 1X SSC/0.1% SDS solution with occasional gentle agitation. After the washes were complete, the slides were rinsed twice in room temperature 1X SSC (plunging 4-6 times) and then in 0.1X SSC. The slides were spun dry at 500 rpm for 5 minutes in a slide box lined with Whatman paper. Arrays were stored in the dark and scanned as soon as possible after they are washed.

3.3 Data collection

Slides were scanned using an Axon GenePix 4000A confocal scanner and fluorescence intensities quantified using GenePix Pro 3.0 software (Axon Instruments, Foster City, CA). A detailed description of this can be found at the website http://www.microarrays.ca/support/Axon%20Scanner.pdf.

The PhotoMultiplier Tube (PMT) detects the photons that are emitted from the laser-excited fluorophores on the microarray. Since increasing the PMT gain increases the sensitivity of the PMT, it was important to use optimal PMT gain. Though it was true that a higher PMT yielded a brighter image, a brighter image was not always a better image. Increasing the PMT increased the noise as well as the signal intensity. If the gain

was too high, the noise increased more than the signal, and the signal-to-noise ratio decreased significantly. When the PMT was set too low, the process of converting photons to electrons was sub-optimal, and the signal-to-noise ratio was low. The best signal-to-noise ratio was achieved when the PMT gain was between 500 and 900. Although it was recommend that the PMTs be balanced so that the fluorescent intensities from both channels were similar, a significant error did not result if the two channels were not perfectly balanced. In reality, the signal intensities from both channels cannot be identical and subsequent software, which implements a normalization factor, corrects this minor imbalance. GenePix Pro 3.0 displays its two single wavelength images at different user-configurable colors, and then constructs a red-green-blue ratio image from them. It includes three different color schemes to display its ratio images: red-green, green-blue, and red-blue. This pseudocolor display provides instant feedback on coarse ratios: with a red-green palette, for example, a ratio of 1 displays as yellow, ratios above 1 show varying shades of red, while ratios of below 1 show varying shades of green. In addition, single wavelength images can be displayed in a number of color palettes: grayscale, monochrome, or two different rainbow palettes.

To measure the intensity of each spot, GenePix Pro 3.0 first assigns a feature indicator to it. Feature indicators are grouped into rows and columns to form a Block. Blocks themselves are grouped into rows and columns to form an array. It can align and analyze the images automatically, for complete hands-free operation. Blocks can be dragged, rotated and deformed; feature-indicators can also be dragged and their diameters fine-tuned, all with simple mouse or keyboard controls. With a mouse-click or keystroke, features can be graphically flagged to mark them as *good, bad or absent.* Flags are exported with all other numerical data. The extracted data is presented in a spreadsheet for rapid sorting. Data such as intensity values, ratios and quality measurements are

exported for analysis with advanced genomic information systems and is fully integrated with web-based genomics databases.

The Feature Viewer function reports the intensity of the pixels contained within the feature-indicator, its associated background intensity level, the precise location of the feature-indicator in the analysis array, and gene IDs or names that have been defined for each feature. Files are saved as GPR (GenePix Results) files. GenePix Pro 3.0 analysis allows extraction of ratio data, production of quality control or interesting genes reports, generation of scatter plots or other custom graphs.

Plots of fold change versus the average intensity were examined to look for abnormalities in single array data. It is common to plot a red versus green scatter plot to examine for problems; however transforming to fold change versus average intensity displayed the data in a more easily viewable form. Plots from all 6 HNSCC cell lines and 28 OSCC samples were found to be of sufficient quality.

3.4 Bioinformatics

The clustering of expression data from the cell lines and tumor samples were performed in 3 separate experiments. The data from the 6 cell lines were clustered by BTSVQ. Data from the initial 20 OSCC underwent unsupervised hierarchical clustering and BTSVQ. Lastly, the full data set from 28 OSCC were combined with the cell line data and clustered using BTSVQ.

<u>Hierarchical clustering of 20 OSCC samples:</u>

Hierarchical clustering was performed on the data set from the 20 OSCC. Each experiment was first normalized by sub array using Normalize Suite. Low intensity spots (<50), ghost spots, and spots with high background to foreground ratios were excluded. Duplicate spot intensities were averaged and the ratio of Cy3/Cy5 calculated. Replicates were combined by first reversing the Cy3/Cy5 ratio in the reciprocal labeling experiments to ensure that the denominator was the tumor sample. Experiments were

combined into a "Project file" that could be loaded into the program Cluster.[82] The normalized mean of ratio values were $\log_2$ transformed and filtered so that only those that were present in at least 80% of the samples and that had at least 2 observations with an absolute value of 2 in $\log_2$ (i.e. at least four fold up or down - regulated) were included in the analysis. Unsupervised average-linkage hierarchical clustering was applied to both genes and samples using a centered correlation similarity matrix. Genes and arrays were median centered. These experiments involved a large number of tumor samples all compared to a common reference sample. The analysis needed to be independent of the amount of a gene present in the reference sample. This was achieved by adjusting the values of each gene to reflect their variation from some property of the series of observed values such as the mean or median. This is what is meant by median centering of genes. Graphical representation of these data was generated with the program TreeView (http://rana.stanford.edu/software)

Clustering of samples with BTSVQ:

For experiments using BTSVQ, raw data from the GPR files were loaded into the software package. It is helpful to *visually* analyze data before applying any clustering or normalization technique. BTSVQ software has the additional benefit of being able to visualize the data in the form of surface plots in three dimensions. Due to outliers or very high ratios in the data, these visualistions often appear skewed. Any clustering method applied on such data will be biased towards the high values. Thus microarray data is generally log-normalized to provide an equal spread between up and down-regulated genes. The transformation removes various types of noise, biases and outliers and often results in a new range of the data that is easier to work with in further analysis. However, the transformation may introduce several distortions and biases, some of which improve the information content, while others may eliminate existing valid patterns. Therefore, to avoid introduction of biases, BTSVQ provides three important normalizations: log,

variance and range. These are applied to the data in sequential order. After each "round" of normalisation, the data was visualized using the surface plots. When satifactory normalization had occurred, the data was then subjected to clustering using the BTSVQ algorithm. This generated a binary tree consisting of clusters or groups of samples at the nodes of the binary tree. The expression profiles of the samples in each group were readily visualized in the form of a SOM, and profiles easily compared.

3.5 Statistical analysis of sample clusters

Sample clusters generated by BTSVQ and hierarchical clustering were subjected to Fisher's exact test, chi squared test, Mann-Witney U test and Log-Rank test using SPSS for Windows statistical software (version 10.1; SPSS Inc., Chicago, IL, USA). Samples clusters were compared to clinical parameters, including sex, age, T and N stage, loco - regional recurrence, 2-year disease free survival and overall survival. Statistical significance was defined as $p \leq 0.05$. Disease free survival was calculated in months and comprised the interval between completion of treatment to the date of first recurrence. Disease specific survival was calculated from completion of treatment to death due to their disease and overall survival from time of completion of treatment to death, whatever the cause. Patients with disease recurrence, or who died of their disease or other causes, were considered censored.

3.6 Results:

3.6.1 Gene expression analysis of 6 HNSCC cell lines:

Expression analysis was first performed on six HNSCC cell lines, using cDNA microarrays containing 19,200 sequences. Gene expression profiles of the cell lines, represented by SOMs, were obtained by BTSVQ analysis. The binary tree is shown in Figure 3. Cell lines were classified in two distinct groups (A and B). Cell lines derived from the same patient (primary and nodal metastases) were clustered together at the first level of the binary tree: UTSCC-60A and 60B in Group A and UTSCC-24A and 24B in Group B. Additionally, cell lines derived from tongue carcinomas clustered together (Group B). The stage of the two tongue tumors were both $T_2N_1M_0$, i.e between 2-4 cm in diameter, with localized neck nodal metastasis (< 3cm) and no evidence of distant metastasis. Although all 4 patients from which the cell lines were derived died of their disease, the two patients in Group B had the worst overall survival, UTSCC – 24A and UTSCC dying of their disease at 9 and 7 months respectively. In Group A, the two primary tumors were both T4, i.e greater than 6 cm in diameter. On the second level of the binary tree, the two cell lines UTSCC-34 (supraglottic larynx) and 60A (tonsil) derived from primary tumors clustered together, while UTSCC-9 (tongue) clustered with UTSCC-24B (neck metastasis).

3.6.2 Gene expression analysis of 20 OSCCs:

The population of 20 samples was representative of the general population with OSCC, with a median age of 61 (range 26-83) and a male predominance of 65%. 75% (15/20) of the patients had TIII-IV tumors, and 65% (13/20) had regional metastases at presentation. None of the patients had distant metastasis.

Figure 4 shows the binary tree generated by BTSVQ for these 20 tumor samples. On the first level of the binary tree, 2 sample clusters were generated. Group 1 contained 12

samples and Group 2 contained 8 samples. The corresponding clinical data for these two groups are shown in Table III and statistical correlations are shown Table IV. In Group 1, 10/12 patients were male, 11/12 tumors were stage TIII-IV and 10/12 tumors were node positive. The median age was 60. 9/12 patients had disease recurrence, 6/12 loco-regional and 3/12 distant. The average time to recurrence was 9 months. 7/12 died of their disease, 1/12 died of other causes and 4/12 had no evidence of disease at last follow-up. The average follow – up in those patients with no evidence of disease was 29 months.

In Group 2, 3/8 patients were male gender, 4/8 were TIII-IV stage and 3/8 samples were node positive. The median age was 63. 5/8 developed loco – regional recurrence at an average time of 5 months. 4/8 died of their disease, 1 was alive with disease at last follow – up and 3/12 showed no evidence of disease at last follow – up. The average follow – up in those patients with no evidence of disease was 30 months. Group 1 and Group 2 correlated with male gender (*P=0.035*), TIII-IV stage (*P=0.035*) and nodal metastasis (*P=0.035*). There were no statistically significant correlations with loco-regional recurrence (*p=0.55*) or 2 year disease free survival (*p=0.42*).

Although Group 1 and 2 did not correlate with prognosis, sample clusters emerged on the lower levels of the binary tree that did show trends towards predicting patient outcome. In Group 1, one patient (005) split from the main cluster on the second level. This patient had a local recurrence and died of disease within 6 months of treatment. On the third level, three patients (34,162,31) split from the main group. All three patients died, two of their disease and one of other causes. On the fourth level, two clusters of four emerged each containing advanced staged tumors (TIII-IV, node positive). The clusters composed of samples with strikingly similar expression profiles. In one group of four (165, 179,111, 29), three patients developed disease recurrence (range 7 – 25 months)

and died of their disease (range 7 - 26 months). In the other group of four (40,58,63,79), 3 patients remain disease free at last follow up (range 12 - 54 months).

Group 2 contained eight patients that partitioned into two groups of four. One of these groups contained all three node positive tumors present in the parent group (86, 96,147,125). These 3 patients all had recurrence (range 2 to 5 months) and died of their disease (5 - 25 months). The other group of 4 contained all non metastatic tumors. Three of these patients had no evidence of disease at last follow up (range 10 - 34 months).

The same dataset was clustered using hierarchical clustering. Following adjustment and filtering of the data set as described in the material and methods section, a representative subset of 2037 genes was used to cluster samples by median centered unsupervised hierarchical clustering. All samples and genes were equally weighted. The dendrogram was viewed using Treeview. A representative portion of this pseudo-colour matrix with the dendrogram showing how the samples were clustered is shown in Figure 5. The samples were split initially into two sample clusters of 10: the red and blue bars below the dendrogram delineate these. In the blue group 8/10 tumors were T III-IV and neck node positive. 8/10 patients developed disease recurrence, 3 distant and 5 locoregional. 6/10 died of their disease, one died of other causes and 3 show no evidence of disease at last follow-up. In the red group, 7/10 tumors were T III-IV and 5/10 neck node positive. 6/10 patients developed disease recurrence, all locoregional. 5/10 died of their disease, 1 was alive with disease and 4 showed no evidence of disease at last follow-up. Correlation with age, sex, T and N stage, recurrence or outcome parameters failed to show any statistical significance. However, of note, 9/10 samples that were present in the blue cluster were also found in Group 1 from the BTSVQ analysis. Similarly, 7/10 samples found in the red cluster were present in Group 2 of the BTSVQ analysis. In the red cluster, samples 33 & 39, 86 & 96 were found together in the sub classification, while sample 29 was classified individually. In the blue cluster, samples 40 & 79, 63 & 58,

179& 165 were also found in the same clusters in the sub classification. These were the same as the clustering found in the lower levels of the binary tree generated by BTSVQ.

3.6.3 Gene expression analysis of 6 HNSCC and 28 OSCC samples:

A further 8 OSCC samples underwent expression profiling using the same methods and materials as described. This data was collated with the 20 OSCC and the 6 HNSCC cell lines. Thus 28 OSCC primary tumor samples and 6 HNSCC cell lines (4 primary and 2 nodal metastasis) were categorized by comparison of global gene expression profiles using cDNA microarrays containing 19,200genes and the BTSVQ data analysis system. Patient demographics, TNM staging and follow-up were available on all the tumor samples and the 4 primary HNSCC cell lines. This study population was representative of the general population with oral/head and neck cancer. The median age was 60, and 75% (24/32) were male. 72% (23/32) of tumors were T3 or T4. 53% (17/32) were node positive, while 46% (15/32) were node negative. None of the patients had distant metastasis.

Figure 6 shows the results of the clustering of the 34 samples using BTSVQ. Initially, seven samples split from the parent group. On the next level, a reasonably balanced split occurred forming two clusters, Group α, containing 16 samples, and Group β containing 11 samples. The demographics, TNM staging and outcome data for these two groups is summarized in Table V. Group α contained one nodal metastasis cell line, which was not included in further analysis. 10/15 patients were male gender with a median age of 59. 6/15 tumors were T I-II and 12/15 were neck node positive. 12/15 tumors recurred. The median follow – up in this group was 12 months, with a 2-year disease free survival of 21% and 2-year disease specific survival of 42%. In Group β, 9/11 were male with a median age of 63. 1/11 was T I-II, and 4/11 neck node positive. 6/11 recurred. The median follow up was 24 months, with a 2-year disease free survival of 55% and 2 year disease specific survival of 64%. Group α contained a statistically higher proportion of

node +ve tumors (p=0.024) compared to Group β. 12/15 patients in Group α recurred as compared to 6/11 in Group β, although this was not significant (p=0.085). Group α also had a worse 2-year disease free and disease specific survival. Disease free survival was statistically significant by log rank analysis of the Kaplan-Meier survival curves (p=0.042) (Figure 7). Although disease specific survival was not significant (0.077), the same divergence was observed. When the same sample population was stratified for disease free survival by N stage, there was no significant correlation by log rank analysis (p=0.085). Classification by BSTVQ was therefore a better predictor of disease free survival than conventional classification by nodal status. On the next level of the binary tree, 2 samples split from Group α and 1 sample from Group β.

3.7 Discussion

HNSCC represents a clinically heterogeneous group of tumors showing different degrees of malignancy. Clinical experience suggests that some carcinomas are destined to remain relatively localized, while others will be widely disseminated at an early stage. Our current staging system classifies these tumors into broad categories based mainly on anatomical considerations. Evidence that further sub typing is possible lies in the fact that staged matched tumors show differences in treatment response and outcome. This coupled with the fact that overall survival of patients with this disease has not changed significantly in the past couple of decades reinforces the need for novel classification systems to improve diagnoses and treatment outcome.

Attempts to find biological markers of tumor progression that may aid classification have been limited by the molecular heterogeneity seen within individual diagnostic categories. No markers identified so far have proven a sufficiently strong indicator of treatment response or prognosis.[125] [141] [126] However, recent studies have used cDNA microarray analysis to classify tumors on global gene expression profiles. These studies fall into two main categories, class discovery and class prediction, of which the later has been the most reported. By identifying genes which best discriminate sample groups, clusters of genes have been identified that predict the known subtype of the disease. These "molecular" subtypes can be used to predict patient outcome or treatment response.

Perou et al characterized variation in gene expression patterns in 65 surgical specimens of human breast tumors from 42 different individuals, using complemantary DNA microarrays representing 8,102 human genes.[99] These patterns provided a distinctive molecular portrait of each tumor. The tumors could be classified into a basal epithelial group, an ERBB2-overexpressing group and a normal breast-like group based on variation of gene expression. Sets of co – expressed genes were identified for which variation in messenger RNA levels could be related to specific features of physiological

function. Subsequently, survival analysis on a sub cohort of patients with locally advanced breast cancer uniformly treated in a prospective study showed significantly different outcomes for the patients belonging to the various groups, including poor prognosis for the basal-like subtype and a significant difference in outcome for two oestrogen receptor-positive groups.[98] In another study on 117 primary breast tumors, DNA microarray analysis and supervised classification was used to identify a gene expression signature strongly predictive of a short interval to distant metastases ("poor prognosis" signature) in young patients without tumor cells in local lymph nodes at diagnosis.[142] The poor prognosis signature consisted of genes regulating cell cycle, invasion, metastases and angiogenesis. In addition, a signature was established that identified tumors of BRCA1 carriers. These findings provided a strategy to select patients who would benefit from adjuvant therapy.

Alizadeh et al proposed that the variability in the natural history of diffuse large B – cell lymphoma (DLBCL), the most common subtype of non-Hodgkin's lymphoma, reflected unrecognized molecular heterogeneity.[100] By microarray analysis, two molecularly distinct forms of DLBCL were identified which had gene expression patterns indicative of different stages of B – cell differentiation. The molecular classification of these tumors on the basis of gene expression revealed previously undetected and clinically significant subtypes of cancer. Furthermore, this classification generated patient groups that differed in survival after treatment with anthracycline-based multi agent chemotherapy regimens. This definition of prognostic groups by gene expression profiling, in combination with clinical parameters, may lead to the recommendation that some patients with DLBCL receive early bone marrow transplantations upon diagnosis.

Bittner et al used several analytical techniques to visualize the overall expression pattern relationships between cutaneous melanoma tumor samples. Using a matrix of Pearson correlation coefficients from a complete pair-wise comparison of all experiments, 31

melanoma experiments were displayed as a hierarchical clustering dendrogram and a multidimensional scaling (MDS) plot. [82] A non-hierarchical clustering system was also employed (termed cluster affinity search technique: CAST) to define experimental clusters.[143] All three analysis systems identified a tight cluster of 19 melanomas. Many genes underlying the classification of this subset were differentially regulated in invasive melanomas that form primitive tubular networks *in vitro*, a feature of some highly aggressive metastatic melanomas.[144] Thus despite the prevailing view that the "taxonomy" of this disease falls in a continuous spectrum lacking discernable entities, classification based on gene expression was possible.[145]

Gene expression profiles have also been extensively investigated in prostate cancer. Dhanasekaran et al used more than 80 cDNA microarrays to assess gene expression in four clinical states of prostate-derived tissues and two distinct reference pools of normal specimens. Benign conditions of the prostate clustered separately from malignant prostate cancer cell lines or tissues, regardless of the reference pool used. Within the prostate cancer cluster, metastatic and clinically localized prostate cancer formed distinct subgroups.

A molecular taxonomy of lung carcinoma was devised from 186 lung tumor samples using oligonucleotide arrays corresponding to 12,600 transcript sequences.[103] Hierarchical clustering defined distinct subclasses of lung adenocarcinoma delineated by high relative expression of neuroendocrine genes and type II pneumocytes genes. Retrospective analysis revealed a less favorable outcome for the adenocarcinomas with neuroendocrine gene expression. The diagnostic potential of expression profiling was emphasized by its ability to discriminate primary lung adenocarcinomas from metastases of extrapulmonary origin. Similarly, global gene expression patterns for 67 human lung tumors representing 56 patients were examined using 24,000-element cDNA microarrays.[104] Subdivision of the tumors based on gene expression patterns faithfully

recapitulated morphological classification of the tumors into squamous, large cell, small cell and adencarcinoma. The adenocarcinoma subgroup was further devided into clusters that correlated with the degree of tumor differentiation as well as patient survival.

Studies have been performed looking at correlations between gene expression profiles of head and neck cancer and clinical parameters or outcome. Using a set of 906 genes, Belbin et al showed it was possible to classify 17 anatomically diverse head and neck tumors into two distinct groups that correlated highly with clinical parameters.[113] Although one group contained more early stage tumors and a lower prevalence of nodal metasases at presentation, this group contained patients with a lower 2 year disease-specific survival (56% vs 100% p=0.057), and overall survival relative to the other group. The molecular classification was a better predictor of disease specific survival than clinical – pathological parameters.

In this study, it was proposed that the phenotypic diversity of head and neck tumors might be accompanied by corresponding diversity in gene expression patterns that could be captured using cDNA microarrays. Systematic investigation of the gene expression patterns might then provide the basis of an improved molecular taxonomy of head and neck cancer. The expression patterns of 6 HNSSC cell lines, 20 OSCC tumors and subsequently a combination of the 6 cell lines and 28 OSCC were identified using microarrays containing 19,200 sequences and two different clustering algorithms.

BTSVQ analysis has been shown to successfully cluster genes across samples, generating useful and visible data organization.[93] Recently, this method was used to cluster gene expression profiles based on the effects of dihydrotestosterone in cell culture.[146] BTSVQ analysis is appropriate for studies using clinically heterogeneous samples, such as HNSCC, where relevant gene expression patterns may not be identified in the first level of the binary tree.[90]

Additionally, it has been shown to be particularly useful in the identification of expression patterns from skewed data or data possessing numerous variable or outliers. Several techniques have been used to visualize this highly multi-dimensional data. The self-organizing map (SOM) algorithm is an efficient tool for the visualization of multidimensional data and has previously been shown to be effective for the exploratory analysis of gene expression data.[147 148] SOMs are neural network algorithms widely used in data analysis and vector quantization. The algorithm is similar to k-means clustering, with the additional constraint that cluster centers are restricted to lie in a two dimensional manifold. SOMs show two main characteristics; they realize a quantization of a high-dimensional space, as other vector quantization techniques such as LBG[149] and k-means, and they exhibit a topological property which allows one to analyze the ordering of centroids. Component planes of SOMs are the planes of Voronoi Tessellations, each representing a specimen in a microarray experiment. Partitive k-means is a splitting hierarchical clustering method. It starts with the whole data set as a single cluster, which is partitioned into disjoint subsets D_1 and D_2, where the inter clusters distance d_{ij} is maximized. The subsets D_1 and D_2 are further subdivided into D_{11} & D_{12} and D_{21} & D_{22}, etc., thus, building a binary tree.

BTSVQ merges the results of SOM (gene space), and partitive k-means (specimen space). The algorithm uses vector quantization and self-organizing capabilities of SOMs in finding significant gene centers in gene space (high dimensionality and large number of clusters), and the effectiveness of k-means in experiment space (medium dimensionality and low number of clusters). The SOM component planes generated by BTSVQ represent a biologically intuitive way of representing gene expression data. Clusters of differentially expressed genes are easily identified by the color-coded scale, which allows easy comparison of expression profiles between samples. Thus the power

of BTSVQ lies in the ability to merge two clustering algorithms allowing visual cross-verification of sample clusters.

The initial part of the study involved the generation, clustering and comparison of gene expression profiles from the 6 head and neck cell lines. Interestingly, cell lines from the same patient were clustered together at the first level of the binary tree; UTSCC- 60 A and B in Group A and UTSCC-24A and 24B in Group B. Given the likely clonal expansion of cells from the primary tumor to the metastatic node, one would expect this observation. Additionally, one would expect cell lines derived from the same anatomical site to be found in the same group. Indeed both primary tongue carcinoma cell lines were found in Group B. Interestingly, these were both derived from the same stage tumor, and the patients had similar survival. The poor prognosis from tongue SCC is related to the frequent incidence of regional metastases. Thus despite aggressive therapy with surgery, radiotherapy and chemotherapy, many patients will succumb to their disease. Despite the primary tumors in Group B having a higher T stage, their overall survival was slightly better. Although T stage is not used clinically as an independent predictor of prognosis, in most tumors of the head and neck high T stage is associated with a higher incidence of metastases and thus poorer outcome. On the second level, UTSSC 34 and 60A clustered together, while in Group B, SCC 9 and 24B clustered together. The fact that the expression profile of the metastases was less similar to its primary tumor in Group A implied that the supraglottic laryngeal and tonsil SCC cell lines had similar gene expression profiles and that the metastasis had undergone further genetic de-regulation resulting in clustering distant from the parent group. The implication from the further sub-classification in Group B, where the primary cell line SCC 9 clustered with the nodal metastasis 24 B, is that metastases in tongue cancer have similar genetic deregulation to the primary site, even though this may not be primary tumor specific. Perou et al showed that in breast cancer, metastasis and primary tumor were as similar in their overall

pattern of gene expression as were repeated sampling of the same primary tumor, suggesting that the molecular program of a primary tumor may generally be retained in its metastases.[99]

HNSCCs are a heterogeneous group presenting as tumors of different anatomical regions such as the oral cavity, the oro-/hypopharynx, and the larynx. These sub entities show different, often contrary, biological and clinical behavior. For example, the potential of forming metastases is very low in laryngeal carcinoma, whereas it is quite common in pharyngeal carcinomas. It has been shown that the various HNSCC sub entities exhibit different genetic alterations, which could explain the different biological behavior of these tumors.[150] The tissue microenvironment can markedly change the gene expression patterns of cancer cells, and therefore their biological behavior, growth ability and potential to metastasize to distant sites [151]. Several studies have been performed that reveal different genetic alteration patterns depending on the anatomical site in HNSCC.[132] [152] Using a combination of tissue microarray and fluorescence in situ hybridization, Freier et al detected a significantly lower rate of ZNF217 amplifications in pharyngeal as compared with oral and laryngeal carcinomas.[153]

Although this pilot study lacked the numbers to reveal statistically relevant associations, the principle of classification of head and neck tumors by gene expression profiles rather than single genetic alterations was demonstrated. Using these samples as proof of principle, it was demonstrated that this type of analysis was able to classify samples into meaningful clusters, as the gene expression profiles were associated with clinically similar tumor sites and clinical characteristics, and thus provided a platform to which the major part of the study could be applied.

In the second part of the study, gene expression profiles from 20 OSCCs were generated and compared. By using a subset of Head and Neck Cancer tumors, the aim was to minimize the genetic differences that may be present in anatomically diverse tumors of

the head and neck region. Additionally, two different clustering algorithms were used to compare and contrast tumor classification, namely BTSVQ and hierarchical clustering. This approach revealed a number of novel observations on the pathobiology of OSCC, most importantly, the presence of molecular subtypes of OSCC correlating with differences in T and N stage.

BTSVQ analysis identified two clusters from the 20 OSCC dataset. Group 1 included 12 OSCC samples and Group 2 was comprised of the other 8 samples. Group 1 included more patients with TIII-TIV category of disease ($p = 0.035$), lymph node metastasis ($p = 0.035$), and also more male patients ($p = 0.035$). There was a slightly higher recurrence (loco-regional and distant) rate and poorer 2-year disease free survival in Group 1, although these outcomes were not statistically significant compared with Group 2.

Although genetic heterogeneity was observed within Groups 1 and 2, tumors with similar clinical characteristics and outcome were classified in distinct branches on subsequent levels of the binary tree. This was visualized by observing the patterns seen in the self-organizing maps (SOMs) generated for each sample. Thus, by further sub-classification of tumor samples, groups with considerable molecular homogeneity were identified. From the 12 patients in Group 1, two groups of four patients were identified in the lowest level of the binary tree. One group contained patients with T3-T4 tumors and lymph node metastases. Three of these patients had disease recurrence and died of their disease within 7-26 months. In this group, patient 29 had no evidence of disease in a follow-up time of 24 months. However, this patient had nodal metastasis at disease presentation. The clustering of this patient's tumor together with the tumors from patients who developed disease recurrence and subsequently died of their disease suggests that this tumor may have the potential to recur. This patient is currently being closely monitored in a follow-up clinic to help in early detection and intervention.

In the other group of four patients, all tumors were T3-T4 and three of four patients have no clinical evidence of disease at last follow up. The fourth patient however, died of disease at 5 months after surgery. Although his tumor profile clustered with those patients with no clinical evidence of disease, visual examination of his tumor profile clearly suggests that it is different from the genetic profiles of patients 40, 58 and 79.

Since there is no single established method to estimate the significance of an observed degree of relationship obtained by cluster prediction, the same dataset was analyzed with a conventional cluster analysis system to investigate whether alternative sample clustering revealed a better correlation with clinical parameters and outcome. The hierarchical clustering algorithm organizes the experimental samples on the basis of overall similarity in their gene expression patterns obtained from filtered data. The clustering relationships obtained are summarized in the dendrogram (Figure 5), in which the pattern and length of the branches reflects the relatedness of the samples. [82] This analysis was based on a subset of 2037 genes generated from adjusting and filtering the initial 19,000 genes that had been screened on the array. Although this is a well established methods of identifying clinically relevant expression profiles, it has been criticized for concentrating on local patterns first, thus loosing the ability to identify patterns present at the global level.[89]

The 20 samples were clustered into 2 groups of 10. These two groups failed to correlate with either T or N stage, or predict patient outcome. Despite this, there were some similarities with the clustering generated from BTSVQ. The observation that 9/10 samples present in the blue group were also present in Group 1 implied that hierarchical clustering was also able to distinguish tumors of advanced stage without any prior biological knowledge of the samples. There were higher numbers of TIII-IV, node positive tumors in the blue group consistent with the BTSVQ clustering in Group 1. These patients had a higher recurrence rate, with more patients developing distant

metastases, implying a more aggressive tumor biology. Similarly, a slightly higher number of patients died of their disease in blue group relative to the red group.

In addition, there were also some similarities between hierarchical clustering and BTSVQ in the sub classification of samples. In Group 2 and the red cluster, samples 33 & 39 were classified together. Both these tumors were TIV N0 derived from the floor of the mouth. Similarly, both patients from which tumor samples 86 & 96 were derived developed early loco-regional recurrences (5 and 3 months respectively) and died of their disease soon after (9 and 5 months respectively). Sample 129, although low stage (TII N0), developed a locoregional recurrence within 2 months of treatment and died within 4 months. This sample was clustered separately in both analyses. In Group 2 and the blue cluster, 40 & 79 were clustered together. Both these tumors were advanced stage (TIII-IV, N1). Despite this, both patients showed no evidence of disease after a substantial follow-up period (24 and 54 months respectively). Similarly, samples 63 and 58 were derived from node positive tumors, and both patients developed early loco regional recurrence. Samples 165 & 179 were derived from advanced stage tongue tumors (TIII-IV) which both developed distant metastasis following treatment (25 and 11 months respectively). Both of these patients died from their disease (26 and 12 months respectively). These results suggested that both unsupervised hierarchical clustering and BTSVQ were capable of distinguishing advanced stage tumors with poorer outcome based on their gene expression profiles and that tumors with similar clinical features and outcome could be classified together. However, BTSVQ appeared to have better delineation of these groups, based on the statistical correlations.

By adding further samples to the initial data set, and combining this with the cell line data, the aim was to identify clusters of samples that not only correlated with clinical parameters, but also had the potential to predict patient outcome. A novel classification system must be capable of providing robust information on treatment response and

outcome. Therefore, if taxonomy based on gene expression profiles is to be clinically useful, it must be better than the current staging system. Correlations of gene expression profiles with conventional predictors of outcome such as T and N stage may not necessarily enhance the classification of tumors. By acknowledging that clinically significant clusters may not initially reveal themselves at the first level of the binary tree, this analysis focused on two sample groups that occurred at the second level. Although these two subgroups did not include all the samples that were originally analyzed, this analysis did highlight that the molecular sub classification of tumors may not be initially apparent from the first dichotomous split, and that clinically interesting groups may occur at subsequent levels, or indeed on different levels. This was seen in the analysis of 20 samples, where the groups that emerged on the lowest level of the binary tree showed trends towards predicting tumor stage and outcome.

Group α contained a lower proportion of TIII-IV tumors but a higher proportion of neck node positive tumors relative to Group β. In addition, more patients developed disease recurrence in Group α, with a statistically lower 2-year disease free survival (p=0.042) and 2 disease specific survival (p=0.077). Thus the molecular classification defined two groups: Group α containing "more aggressive" tumors and Group β containing "less aggressive" tumors. Interestingly, when T or N stage was used as independent predictors of disease free survival, there was no statistically significant correlation (p=0.12, p=0.085 respectively). Therefore, the classification of tumors based on gene expression profiles was better predictor of outcome than conventional clinical parameters. The implication from this classification is that T stage has a less significant impact relative to N stage on the likelihood of developing disease recurrence, on 2-year disease free and 2-year disease specific survival. In clinical practice, this observation is often seen in head and neck cancers, whereby large primary tumors (> TIII) have no regional metastases, and small primary tumors (<TII) are associated with large neck lymphadenopathy. The

presence of lymph node metastasis has been shown to have a detrimental affect on outcome.[154] Interestingly, on the third level of the binary tree, only two patients were split from Group α and only 1 patient from Group β, implying that these two major groups were genetically homogeneous and that the analyses was reluctant to partition the groups further.

Recent studies involving gene expression profiling of clinical specimens have had a profound impact on cancer research. In many examples, correlations have been made between the expression levels of a gene or set of genes and clinically relevant sub classifications of tumor subtypes. These results have compounded expectations that true molecular classification and sub staging of multiple tumor types may be possible, leading to measurable improvements in prognosis and patient management. Overall, these results suggest a number of implications for the classification and staging of head and neck tumors. Most prominent is that a true molecular staging system, built on either the current system or constructed anew, has the potential to additionally refine diagnosis, prognosis, and management for patients with this lethal disease.

CHAPTER 4: Identification and validation of differentially expressed genes.

4.1 Introduction:

Gene expression profiling allowed a molecular classification of OSCC and HNSCC cell lines. These "molecular maps" were correlated with clinical parameters and outcome. The next challenge was to identify the genes that were the "best predictors" of these statistical correlations. Since hierarchical clustering of 20 OSCC dataset had not identified any statistically significant correlations with clinical parameters or outcome, the BTSVQ method for identification of genes was used. BTSVQ selects a unique set of genes that best discriminates each sample cluster positioned at each node of the binary tree. Selection is based on the gene quantization error (QE). This is the probability of the gene having a similar expression value across all samples in that cluster, thus providing a measure of the ability of a gene to define a cluster of samples. Genes were ranked from "best predictive" to "worst predictive". The lower the QE score, the better predictor of the sample cluster. Since Group 1 correlated with high T and N stage, genes in this sample cluster may represent potential biomarkers of advanced stage disease and thus the genes that best represented this group were chosen for further study. Additionally, since the classification of 34 samples resulted in clusters that were able to predict disease free survival, genes were also identified that discriminated Group α and β.

Furthermore, a subset of the genes representing Group 1 were selected for validation by quantitative real time PCR. This allowed the comparison of individual expression levels obtained from an independent technique with those obtained from the microarray experiments. The mean levels of expression of the validated genes were compared and their power to predict sample clusters investigated. Correlation co-efficients between these genes in each group allowed further interrogation of the hypothesis that they were

the best predictors of the sample classification generated by the full microarray dataset. Additionally, individual gene expression ratios from the 20 sample microarray dataset were mapped to respective chromosomal localization, and the localizations of the 6 validated genes compared with areas of chromosomal gain representing transcriptional over activity.

4.2 Methods:

4.2.1 Gene identification:

Genes that discriminated Group 1 were identified based on their QE score (Table VI). Quantization error does not replace statistical significance, rather, is used to rank-order genes based on their potential to cluster samples. Thus QE score provided an accurate method of excluding genes with variable expression across samples. Genes were selected with a QE score <0.1. This was essentially an arbitrary cut-off point, but served to reduce the gene set to a manageable number for investigation. This number was large enough to be robust against noise, and small enough to be readily applied in a clinical setting. Gene sequences were mapped to Unigene clusters and functions explored to evaluate their potential relationships with malignancy. The same cut – off QE score of <0.1 was used to identify genes that were the best predictors of Group α and β (Table VII).

4.2.2 Validation by Quantitative Real-Time RT-PCR:

Based on the in - silico analysis of protein products and mapping information, six out of the 23 genes from Group 1 were selected for expression validation by quantitative real time PCR. Real-time chemistries allow for the detection of PCR amplification during the early phases of the reaction. Measuring the kinetics of the reaction in the early phases of PCR provides a distinct advantage over traditional PCR detection. Traditional methods use agarose gels for detection of PCR amplification at the final phase or end-point of the PCR reaction. Results are based on size discrimination, which may not be precise and vary from sample to sample. While gels may not be able to resolve this variability in yield, real-time PCR is sensitive enough to detect these changes. Agarose gel resolution is very poor, about 10 fold, while real time PCR can detect as little as a two-fold change. Theoretically, there is a quantitative relationship between the amount of starting target

sample and amount of PCR product at any given cycle number. Real-time PCR detects the accumulation of amplicon during the reaction. The data is then measured at the exponential phase of the PCR reaction, the optimal point for detection of product.

The genes selected for validation were *CIDEB*, *PRKAR2A*, *ANP32A*, *GALNT6*, *SMARCC2* and *CLDN1*. The choice of samples to perform this validation on was limited because of the quantity of RNA available from each specimen. This meant that validation could only be performed on a subset of the total sample size.

The ABI Prism 7700 Sequence Detection System (PE Applied Biosystems) was used for quantitative assessment of expression level of the genes selected for validation. This was done using SYBR Green I dye, which allows the quantitative detection of specific product accumulation during each PCR cycle.[155] SYBR Green is a dye that binds the minor groove of double stranded DNA. As more double stranded amplicons are produced, SYBR Green dye signal increases. A CCD camera collects the fluorescence emission from the amplified DNA and the data are quantified and analyzed using the Sequence Detection System software v. 1.7 (PE Applied Biosystems).

4.2.3 Genes and Primers:

cDNA sequences for each gene were selected from the international published databases (http://www.ncbi.nlm.nih.gov/). Primers were designed for optimal hybridization kinetics using Primer Express software (version 1.5, PE-Applied Biosystems, Foster City, CA, USA) (Table VIII). A full explanation of the details of primer design using this software can be found at the website http://www.appliedbiosystems.com/support/tutorials/.

4.2.4 PCR Amplification:

Reaction mixtures contained cDNA reverse transcribed from 2μg of total RNA from each tumor sample or from Human Universal RNA (Stratagene, Vancouver, BC,

Canada), 20μM of each primer and 12.5μl of 2X SYBR Green PCR Master Mix (PE *Applied Biosystems*), which includes the SYBR Green I fluorescent dye, 0.5 units of AmpErase uracyl-N-glycosylase (UNG) enzyme, 1.25 units of Ampli-Taq Gold DNA polymerase, and 200μM of dideoxynucleotides. Amplification conditions were the same for all primer sets. Thermal cycling conditions were 50°C for 2 min (for UNG enzyme activity), 95°C for 10 min, and 40 cycles at 95°C for 15s followed by 60°C for 1 min. Each assay included a human universal cDNA reverse transcribed from Human Universal RNA (Stratagene, Vancouver, BC, Canada) and a non-template control. Experiments were performed in triplicate for each sample in the same reaction and repeated when a coefficient of variation higher than 5% was observed. Target genes and the reference gene *GAPDH* were amplified in the same reaction.

4.2.5 Quantitative Real-Time RT-PCR Data Analysis:

Quantification of the expression of the target gene in unknown tumor samples is accomplished by measuring the fractional cycle number (Ct) at which the amount of expression reaches a fixed threshold and is directly related to the amount of product. The threshold line is set in the exponential phase of the amplification for the most accurate reading. The precise amount of total cDNA added to each reaction is difficult to assess. Therefore, quantified products of the housekeeping gene *GAPDH*, which maps to 12p13, were quantified as an internal control. The *GAPDH* gene is frequently used in quantitative studies using DNA or RNA.[156] This method normalizes the amount and the quality of the amplified targets.[157] The relative quantification is given by the Ct values, determined for triplicate reactions for test and reference samples for each target and for the internal control gene (*GAPDH*). Triplicate Ct values were averaged and the *GAPDH* Ct subtracted to obtain ΔCt [ΔCt = Ct (target gene) – Ct (*GAPDH* gene)]. ΔCt values were calculated for each tumor and reference sample. Relative expression level was determined as $2^{-\Delta\Delta Ct}$, where ΔΔCt = ΔCt (target sample) - ΔCt (reference sample). For the

reference sample, ΔΔCt equals zero, and 2^0 equals one, so the fold change in the reference sample equals one, by definition. For the unknown samples, evaluation of $2^{-\Delta\Delta Ct}$ indicates the fold change in gene expression relative to the reference sample.[158] The values were expressed as N-fold differences in target gene expression. Representative plots of the RT-PCR results for GAPDH and the validated gene CLDN 1 are shown in Figure 8.

4.2.6 Statistical Analysis of Quantitative Real-Time RT-PCR Results:

Using SPSS for Windows statistical software (version 10.1; SPSS Inc., Chicago, IL, USA), relative expression levels obtained from quantitative real-time RT-PCR for the six validated genes were subjected to the t-test to identify the differences in expression levels of these six genes between Groups 1 and 2 samples. As multiple comparisons were performed on the real-time PCR data from these six genes, Bonferroni correction was used to adjust the level of statistical significance to $p \leq 0.0083$. This correction was applied to all real-time PCR data analysis. Other studies have used an adjustment based on Dubey's approach to preserve overall significance in multidimensional data.[94]

In order to determine the predictive value of the six validated genes, an unsupervised BTSVQ analysis using the quantitative real-time RT-PCR data was performed. The clustering of samples was compared to the binary tree generated from the full microarray data set.

To identify correlations in levels of expression between genes in each sample cluster, a correlation matrix was generated using Matlab R12 software (MathWorks Inc., Natick, MA, USA) to display the correlation coefficients for the six validated genes in the two clinically different groups of samples. The pseudo-color presentation of correlation coefficients represent a visually intuitive way of comparing sample groups.

An additional feature of Normalise Suite is the *Profiler* software. Profiler uses updated chromosome localization information to generate composite chromosome plots of

normalized fluorescence intentsity ratios from selected samples, generating images analogous to karyotype profiles from chromosome CGH. Only results for cDNA sequences with known cytoband and megabase positions are plotted. To identify up-to-date cytoband localizations for the cDNA's housed on the 19k chip, custom software was used to parse NCBI UniGene and MapViewer databases. Gene expression ratios were thus mapped to their respective chromosomal localizations using Profiler. The chromosome locations of the six genes selected for validation were investigated to see if they lay in areas of transcriptional over-expression.

4.3 Results:

Gene identification:

BTSVQ analysis identified a subset of 23 differentially expressed genes with a QE < 0.1 in Group 1 samples. These genes are shown in Table VI, along with their respective Unigene ID, cytogenetic location and function. From these 23 genes, 11 were Expressed Sequence Tags (ESTs), hypothetical proteins and KIAA sequences with unknown function, and 12 were known genes. Of the known genes, a number have known or potentially interesting relationships with putative mechanisms of malignancy and metastasis. Since these genes were the best predictors of Group 1, these genes may represent putative markers of advanced stage tumors and regional metastasis.

Similarly, BTSVQ analysis identified a subset of 19 differentially expressed genes with a quantization error (QE) < 0.1 in Group α samples, and 5 genes in Group β. Gene accession number, Unigene ID and function are shown in Table VII. In the cluster of 19 genes, 13 mapped to unigene clusters, 9 of which were predicted proteins found as EST's or hypothetical proteins with no annotated function. In the cluster of 5 genes, 3 mapped to unigene clusters with annotated functions. Since Group α and β correlated with disease free survival these genes may represent potential biomarkers of outcome for head and neck cancer patients.

Validation:

6 out of the 23 genes with QE scores < 0.1 that discriminated Group 1 were selected for validation by quantitative real-time RT-PCR. This selection was based on the in-silico analysis of protein function, mapping information and association with putative mechanisms of malignancy. These genes were *CIDEB*, *PRKAR2A*, *ANP32A*, *SMARCC2*, *CLDN1* and *GALNT6*. Expression of these genes was tested in 14/20 OSCC samples

used for microarray analysis. The other cases were not analyzed because of insufficient RNA. Of these 14 cases, nine cases (05, 29, 31, 40, 58, 79, 111, 165, 179) were from Group 1 and five cases (33, 39, 96, 120, 147) were from Group 2.

Relative transcript levels of all genes obtained by quantitative real-time RT-PCR were consistent with the microarray data. Samples from Group 1 had higher expression levels of these genes than samples from Group 2. Group 2 samples did not show differences in the transcript levels of these six genes; *i.e.*, these samples showed transcript levels similar to those of human universal RNA. The relative transcript levels of the six genes in samples from Groups 1 and 2, as detected by quantitative real-time RT-PCR, are shown in Table IX and Figure 9.

A comparison between Groups 1 and 2 of the relative expression levels of the six genes obtained from quantitative RT-PCR was performed using the t-test analysis. A statistically significant correlation was observed between *CLDN1* over-expression and the cluster of samples (Group 1 *vs*. Group 2; $p = 0.007$) after Bonferroni correction. The other five genes were also all over-expressed in Group 1 samples, as compared to Group 2 samples. However, over-expression of these five genes was not statistically significantly correlated with the clustering of samples in Groups 1 and 2 after Bonferroni correction (level of significance adjusted to $p \leq 0.0083$). These data are presented in Table X.

BTSVQ analysis of the relative expression levels determined by quantitative RT-PCR was applied to cluster the 14 OSCC samples. These results are shown in Figure 10A. Interestingly, Group 1 and Group 2 samples split at the first level of the binary tree, the same result previously obtained by BTSVQ analysis of the entire microarray data set. These results suggested that the combination of the validated expression levels of these six genes was able to predict the clustering of the samples into two groups that correlated with advanced stage disease.

Pseudo-color correlation matrices were constructed in order to further examine the hypothesis that the six genes could play a role in separating Group 1 and 2 samples. These results are shown in Figure 10B. The *color map* corresponds to the scale of correlation coefficients: non-correlated data show as *light blue*, negative correlation *dark blue*, and positive correlation ranging from *yellow* to *red*. The diagonal of the symmetric correlation matrix represents self-correlation (*r=1.0, dark red*). Interestingly, while Group 1 showed high correlation between *CIDEB* and *PRKAR2A* (*r=0.71*), Group 2 showed a negative correlation (*r=-0.2*) between these two genes. Similarly, Group 2 showed high positive correlation of *PRKAR2A* and *GALNT6* (*r=0.71*), *ANP32A* and *CIDEB* (*r=0.52*), and *PRKAR2A* and *SMARCC2* (*r=0.53*), while Group 1 showed negative correlations (*r=-0.44*; *r=-0.2*; *r=-0.47*, respectively).

The results of the ideograms generated by the Profiler software for 4 chromosomes are shown in Figure 11. The normalized fluorescence intensity ratios per project (sample) are shown in yellow, and their average is plotted in red. The black arrows pinpoint the localizations of the six genes selected for validation. Since the microarray ratios were derived from expression of RNA, the peaks in the traces map to areas of transcriptional over – expression. Interestingly the validated genes were located in chromosomal loci consistent with transcriptional gain. Additionally, genes *PRKAR2A* and *CLDN1* mapped to distinct relevant regions on chromosome 3: 3p21.3 and 3q28-q29 respectively, while *GALNT6* and *SMARCC2* genes mapped to loci approximately 4Mb on chromosome 12: 12q13.13 and 12q13-q14, respectively.

4.4 Discussion

Several studies have used cDNA microarray analysis to identify differentially expressed genes involved in development and/or progression in head and neck cancer. Belbin identified 375 genes which devided 17 patients with head and neck tumors into two clinically distinct subgroups.[113] The selection of genes was based on the ranking of genes that best discriminated the two subgroups using an elemental t test. The genes originated from diverse functional categories and included genes already associated with neoplastic disease states. Of note, genes involved in transport (SLC7A8), RNA processing and translation (ribosomal proteins S6 and S19) were up regulated in one group. Additionally, genes coding for metabolic enzymes (catalase and xanthine oxidase) and genes whose expression is known to be induced by TGF-β (TGF-β inducible early response 2, TGF-β activated kinase 1b) were also up regulated.

Using a subtractive library from two HNSCC and six normal tissues, Villaret et al used a custom built chip to identify 13 independent genes in 16 anatomically heterogeneous tumors compared with 22 normal tissues.[114] 9 of these genes were previously known: keratins K6 and K16, laminin-5, plakophilin-1, matrix metalloproteinase-2, vascular endothelial growth factor, connexin 26, 14-3.3 sigma, and CaN19. The cytokeratin genes K6 and K16 were the most commonly overexpressed genes, a finding consistent with head and neck literature.[159] Partial sequencing of four gene elements failed to show any homology with a search of the GenBank database; thus they were labeled as novel. However, only one of these genes showed significant difference between the average expression of the tumor and the control groups.

Leethanakul *et al* identified a number of under- and over-expressed genes in matched tumor and normal tissues in five HNSCC patients using lazer capture microdissection to procure pure samples and arrays containing 588 known cancer genes.[116] Of note there was a significant decrease in the expression of cytokeratins (2-20 fold) in the tumor

samples compared with normal, most likely representing loss of differentiation in the tumor cells. Additionally, there was a clear increase in cyclin – D1, metalloproteases and many growth and angiogenic factors including TGFα, TGFβ , EGF, PDGF A chain and B chain, different FGF isoforms, HGF and VEGF-C. This supported the conclusion that this tumor type secretes factors that are likely to induce epithelial cell gowth in an autocrine fashion in addition to promoting the growth of stromal cells and the process of neovascularisation. Recently, these authors used bioinformatic tools to retrieve the available sequence information from the Head and Neck Cancer Genome Anatomy Project (HN-CGAP) database. They compared this information with their data generated from cDNA libraries of normal tissue and oral carcinomas of different stages.[160] 55 known genes were identified from this sequence analysis. Of particular interest was the sequence match to the monocyte chemotactic protein 3-precursor gene (MCP3), which was readily detected in all three cDNA libraries from tumor tissues. MCP3 is a chemokine known to induce the production of gelatinase B and chemotaxis of moncytes. MCP3 is also known to be produced by tumor cells and its expression in HNSCC may play a role in tumor progression.[161] To address which of the novel genes were frequently expressed in HNSCC, their DNA was used to construct an oral-cancer-specific microarray, which was used to hybridise α-^{33} P dCTP labeled cDNA derived from five HNSCC patient sets.[162] Initial assessment demonstrated 10 clones to be highly expressed (> 2 fold) in the normal squamous epithelium, while in three of the five patient sets, 14 were highly expressed in the malignant counterpart, thus suggesting that a subset of these newly discovered transcripts might be expressed in this tumor type. [115] Similar to these experiments, Alevios et al performed large scale gene expression profiling on lazer capture microdissected tumor and normal oral epithelial cells analysed on high-density oligonucleotide microarrays.[115] About 600 genes were found to be oral cancer associated, including oncogenes, tumor suppressors, transcription factors, xenobiotic enzymes,

metastatic proteins, differentiation markers, and genes that have not been implicated in oral cancer.

Al Moustafa et al used four matched primary normal epithelial and HNSCC cell lines to identify 91 up –regulated genes (ratios 2.5 – 200) and 122 down-regulated genes (ratio – 200 to 2.5) from an Affymetrix array housing 12 530 human genes.[131] Consistent with Leethankul et al., the expression of cytokeratins (13 ,15 ,17, 18), cyclin D2, β-catenin, TGF –β, VEGF, FGF and Wnt were altered in cancer in comparison with normal cells. The expression of 9 selected genes from the cell-cell adhesion and motility group were investigated by Western and/or RT-PCR. As expected from the array analysis, Wnt-5a, N-cadherin and fibronectin were found to be up-regulated; whereas claudin-7, E-cadherin, α-catenin, β-catenin, γ-catenin expression were down-regulated. For the first time, they confirmed under expression of claudin-7 and connexin 31.1 by microarray, western blotting and/or RT-PCR in HNSCC. Claudin 7 is a four transmembrane domain-containing protein and is a member of a recently identified family of proteins, the claudins.[163] The main function of claudins appears to be critical components of tight junctions, and therefore could have important cell adhesion functions in HNSCC. Squire *et al*[117] identified and correlated recurrent gains of 3q and 8q with differentially expressed genes located in these chromosomal regions, as detected by comparative genomic hybridization (CGH), spectral karyotyping (SKY) and expression array analysis, in seven head and neck cancer cell lines and five primary tongue tumors.

More recently, high-density DNA microarrays and quantitative real-time RT-PCR and have identified 45 deregulated genes in oral cancer.[164] A set of well-characterized statistical tools which included Wilks' lambda score, error rate estimated from leave-one out cross-validation (LOOCV) and Fisher Discriminant Analysis (FDA) were used to generate a robust classifier using the identified discriminatory genes. Of the 45 genes

identified, six have been previously implicated in oral cancer, and two are uncharacterized clones.

BTSVQ identifies sample clusters based on the analysis of the full data set. This allows the identification of clusters based on unfiltered data, thus allowing true global gene expression profiling. Genes of significance that represent each cluster are ranked from "best predictive" to worst predictive based on their quantization error. This is best described as the likelihood of the gene having a similar expression level across all samples in that cluster. Because the data has not been previously filtered, this method may identify genes with only a small deviation in transcript level from normal. However, other studies have shown that genes that are not necessarily expressed over or under a certain threshold level may still be important in disease pathogenesis. Thus selection based on QE score rather than threshold level may be a more powerful tool in the identification of biologically important genes.

BTSVQ analysis identified 23 differentially expressed genes with QE scores <0.1 in Group 1. Of these 23 genes, six genes were selected for validation using quantitative real-time RT-PCR. The criteria for selection of these genes for validation were based on known mapping information and function, and their putative roles in disease. In order to assess the differential expression of these genes in the two clusters and to confirm the microarray results, samples from Groups 1 and 2 from which enough RNA was available were included in the validation study. The expression levels of six genes were validated using quantitative real-time RT-PCR.

CIDE-B (cell death-inducing DFFA-like effector B), encodes a cell-death inducing protein, and has a role in apoptosis induced by DNA damage.[165] [166] Homology is shared with the N-terminal region of DFF (DNA fragmentation factor). The exact nature of the molecular mechanisms of CIDE-B-induced apoptosis is unclear. The CIDE-B protein is localized in mitochondria and forms homodimers and heterodimers with other family

members. Serial deletion analyses suggest that the mitochondria localization signal and dimerization interface are overlapped and localized to the 30 amino acid residues at the C-terminal region of CIDE-B. Mitochondria localization and dimerization are both required for CIDE-B-induced apoptosis. This gene has been found to be down-regulated in mucoepidermoid carcinoma.[167]

GALNT6, encodes a member of the UDP-N-acetyl-alpha-D-galactosamine polypeptide:N-acetylgalactosaminyltransferase family of enzymes.[168] The levels of mRNA expression of three transferases from the same family were quantified in human adenocarcinoma cell lines from pancreas, colon, stomach, and breast.[169] Two of the GalNAc-transferases, GalNAc-T1 and GalNAc-T2, were expressed constitutively and at low levels in most or all cell lines examined. A third GalNAc-transferase, GalNAc-T3, was differentially expressed. Well-differentiated adenocarcinoma cell lines expressed high levels and moderately differentiated cell lines expressed lower levels of GalNAc-T3. Thus glycosylation in tumor cell lines may be regulated in part by differential expression of GalNAc-transferases and gene expression may be a molecular indicator of differentiated adenocarcinoma.

The SWI/SNF-related, matrix-associated, actin-dependent regulators of chromatin (SMARC), also called BRG1-associated factors, are components of human SWI/SNF-like chromatin-remodeling protein complexes and are involved in transcription regulation and coactivation. Human SMARC genes have been mapped to regions on four different human chromosomes, SMARCC1 to 3p23-p21, SMARCC2 to 12q13-q14, SMARCD1 to 12q13-q14, SMARCD2 to 17q23-q24, and SMARCD3 to 7q35-q36. SMARCC1, SMARCC2, and SMARCD1 are assigned to chromosomal regions that are frequently involved in somatic rearrangements in human cancers.[170]

ANP32A (acid (leucine-rich) nuclear phosphoprotein 32 family, member A) is involved in signal transduction.[171] Interestingly, a sequence identity study showed that the

ANP32A gene is identical to the protein named SET, which is encoded by the gene *SET* (SET translocation myeloid leukemia-associated). This gene maps to 9q34 and was observed to be fused to the putative oncogene *CAN* in a patient with acute undifferentiated leukemia.[172]

PRKAR2A encodes for the regulatory subunit RII alpha of cAMP-dependent protein kinase.[173] Other protein kinases have been shown to regulate human head and neck squamous cell carcinoma motility, adherence, and cytoskeletal organization.[174]

Finally, *CLDN1* (Claudin 1), encodes an integral protein component of tight junctions and is involved in controlling cell-to-cell adhesion.[163] *CLDN1* is expressed in cell types such as epithelia, endothelia, and perineural cells. The expression of this protein has also been observed exclusively in perineuriomas, aiding in the histological classification of these tumors.[175] Increased expression of *CLDN1* has been reported in primary colorectal cancers when compared to adjacent non-cancerous mucosa. Furthermore, IHC staining showed a high expression of this protein in cancer cells, suggesting a role of this gene in colorectal tumorigenesis.[176] *CLDN1* showed decreased expression in breast tumor cell lines, however, no mutations in the promoter or specific exons were found to support its under-expression and its role in the breast tumorigenesis process.[177] Interestingly, other members of the claudin gene family have also been reported to be over- or under-expressed in cDNA microarray studies of head and neck tumors.[131]

Quantitative real-time RT-PCR results confirmed the over-expression of the above six genes in all the tumor samples from Group 1 that were used for validation. Group 2 samples showed no differences in expression, as compared to human universal RNA. In the nine tumor samples from Group 1, five of the six genes had increased relative transcript levels in comparison to human universal RNA. The *PRKAR2A* gene only showed a slight increase in gene expression. However, studies have suggested that genes

showing a relatively small change in transcription may still be functionally important in disease etiology and/or progression.[94]

The relative expression levels of the six genes, as detected by quantitative real-time RT-PCR, were analyzed using the t-test. All six genes were significantly correlated with sample clusters using a conventional p value of ≤ 0.05; however, the expression levels of five of these genes failed to reach significance after adjusting the level of significance to $p \leq 0.0083$ using Bonferroni correction. The Bonferroni correction is a mathematical correction utilized to reduce false positive results derived from statistical analyses where multiple comparisons are performed on the same data set. This correction has been used to adjust the level of significance in other expression studies.[178] Therefore, after Bonferroni correction, *CLDN1* was the only over-expressed gene significantly correlated with the sample cluster containing more advanced stage tumors ($p = 0.007$).

In addition, the relative expression levels for the six validated genes detected by quantitative real-time RT-PCR were analyzed using BTSVQ. The clustering of samples was identical to the clustering obtained previously with the full data i.e the 9 samples from Group 1 clustered together and the 5 samples from Group 2 clustered together. Furthermore, evidence that these genes were the best predictors of the sample clustering came from the correlation matrix generated to identify patterns of correlation between genes within each group. Similar correlation matrixes have been used in other studies to identify relationships between genes.[179-181] The first observation was that the pseudo-colors generated to visualize the correlation co-efficients were significantly different, implying that correlations between paired genes could also be used to dichotomise the two groups. The second observation that the correlation co-efficients were positively and negatively correlated between groups further reinforces the case for these genes ability to predict sample classification.

The ideograms generated from the Profiler software demonstrated that the 6 genes were located on cytogenetic bands with high transcriptional activity. The genes *PRKAR2A* and *CLDN1* map to distinct relevant regions on chromosome 3: 3p21.3 and 3q28-q29 respectively. CGH studies indicate that these two regions are amplified in head and neck cancer.[182-184] Gain of 3q is one of the earliest genetic markers for invasion and metastasis and correlates with poor prognosis.[183] This region (3q26.1-29) is also reported as the smallest recurrent chromosomal region of high-level amplification in OSCC using CGH.[132]

The *GALNT6* and *SMARCC2* genes, mapping to 12q13.13 and 12q13-q14, respectively, are approximately 4Mb apart. Gains in this chromosomal region are correlated with a poor prognosis and metastasis in lung[185] and head and neck cancer patients.[186] In addition, recurrent DNA copy number loss of 12q with a minimal common overlapping region at 12q12-q13 was reported in adenoid cystic carcinoma.[187] The concomitant differential expression of these genes suggests that this region may have an important role in the prediction of prognosis.

BTSVQ identified several genes of interest that were the best predictors of Group α and β obtained from the classification of 34 samples. Prosaposin is a multifunctional protein that has been reported to be secreted by breast cancer cells[188], induce extracellular signal-regulated kinases and sphingosine kinase activity, increase DNA synthesis, and prevent cell apoptosis.[188 189] Studies have suggested its putative role in eliminating barriers to tumor metastasis by facilitating hydrolysis of membrane glycolipids.[188] CD40 is a M_r 45,000 to 50,000-glycoprotein member of the tumor necrosis factor receptor superfamily expressed on the surface of a variety of cells. Podner et al. showed CD40 expression on the cell surface of 7 HNSCC cell lines and provided putative evidence for its function in cell growth regulation.[190] In OSCC, loss of polarised expression of CD40L and maintained expression of CD40 might be involved in tumorgenesis and immune evasion

[191], and expression in metastatic lung cancer has raised the possibility of CD40 as a prognostic marker and an indicator of advanced disease.[192] In a Phase I study, recombinant CD40 ligand showed encouraging antitumoural activity in a patient with advanced laryngeal cancer.[193] Up-regulation of Bcl-2 expression in endothelial cells that constitute tumor microvessels has been shown to enhance intratumoral microvascular survival and density, and accelerate tumor growth.[194] Neoplastic conversion of human urothelial cells has been shown in vitro by overexpression of H_2O_2-generating peroxisomal fatty acyl CoA oxidase[195]. An increase in the expression of cyclin G1 during sequential development of liver cancer has been observed, and therefore may play an important role in tumor progression in oral cancer[196]

Many clinical studies have correlated alterations in expression of individual genes with clinico – pathological criteria and outcome, often with contradictory results.[141 197-203] Surprisingly none of these more "classical" markers were present in the set of marker genes identified by BTSVQ. This could be due to the fact that gene expression was determined at the level of transcription in microarray experiments, whereas most of these studies measured protein levels. However, it is more likely that these genes in isolation have limited predictive power, which highlight an approach based on many genes. Additionally, the absence of genes already associated with head and neck cancer was perhaps not surprising: (a) some genes, such as cyclin D1 was not on the array used in this study; and (b) because this study specifically addressed differences in gene expression between the primary tumors themselves, one would not necessarily expect to see genes identified previously by comparing tumor cells with corresponding "normal" cells. Thus, identification of a cluster of genes with distinct biological functions, whose combined expression profile correlates with disease, may prove a more powerful tool in identifying clinically relevant biomarkers, as compared to single gene studies. The mechanisms by which alterations in the activities of these genes may influence tumor

growth, development and metastasis may include over-expression of normal gene products, gene amplification or mutation.

4.5 Conclusions

Novel classification systems of head and neck cancer are crucial to improve patient-specific treatment strategies and outcomes. These findings were consistent with recent data that link classification based on a gene or set of genes to clinically relevant sub-classifications of tumors.[88 94 113 116] Further investigation of these genes is required to ascertain their validity and potential use as diagnostic markers or markers of disease progression in head and neck cancer. Studies such as these will allow identification of gene clusters that are relevant to the diagnosis and prognosis of patients with head and neck cancer, and to design disease-specific gene arrays. The use of such "onco-chips" may ultimately result in improvements in patient outcome due to the development of more tailored therapeutic strategies for the treatment of head and neck cancer.

Acknowledgements

This work was supported by NCIC grant (013220) to Dr. Suzanne Kamel-Reid, NSERC grant (203833), IRIS grant to Dr. Igor Jurisica, and The Sir Harry Morton Travelling Fellowship from the Royal College Surgeons of England.

Personal thanks to Nigel Beasley for his help with the application for the grant fom the Colleges of Surgeons, and to Patricia Reis for teaching and guidance with RT-PCR.

REFERENCES:

1. Greenlee RT, Hill-Harmon MB, Murray T, Thun M. Cancer statistics, 2001. *CA Cancer J Clin* 2001;51(1):15-36.
2. Boyle P, Macfarlane GJ, Zheng T, Maisonneuve P, Evstifeeva T, Scully C. Recent advances in epidemiology of head and neck cancer. *Curr Opin Oncol* 1992;4(3):471-7.
3. Koch WM, Lango M, Sewell D, Zahurak M, Sidransky D. Head and neck cancer in nonsmokers: a distinct clinical and molecular entity. *Laryngoscope* 1999;109(10):1544-51.
4. Sankaranarayanan R, Masuyer E, Swaminathan R, Ferlay J, Whelan S. Head and neck cancer: a global perspective on epidemiology and prognosis. *Anticancer Res* 1998;18(6B):4779-86.
5. Do KA, Johnson MM, Doherty DA, Lee JJ, Wu XF, Dong Q, et al. Second primary tumors in patients with upper aerodigestive tract cancers: joint effects of smoking and alcohol (United States). *Cancer Causes Control* 2003;14(2):131-8.
6. Allison PJ. Factors associated with smoking and alcohol consumption following treatment for head and neck cancer. *Oral Oncol* 2001;37(6):513-20.
7. Brennan JA, Boyle JO, Koch WM, Goodman SN, Hruban RH, Eby YJ, et al. Association between cigarette smoking and mutation of the p53 gene in squamous-cell carcinoma of the head and neck. *N Engl J Med* 1995;332(11):712-7.
8. Somers KD, Merrick MA, Lopez ME, Incognito LS, Schechter GL, Casey G. Frequent p53 mutations in head and neck cancer. *Cancer Res* 1992;52(21):5997-6000.
9. Gronbaek M, Becker U, Johansen D, Tonnesen H, Jensen G, Sorensen TI. Population based cohort study of the association between alcohol intake and cancer of the upper digestive tract. *Bmj* 1998;317(7162):844-7.
10. Oude Ophuis MB, Roelofs HM, van den Brandt PA, Peters WH, Manni JJ. Polymorphisms of the glutathione S-transferase P1 gene and head and neck cancer susceptibility. *Head Neck* 2003;25(1):37-43.
11. Henderson CJ, Smith AG, Ure J, Brown K, Bacon EJ, Wolf CR. Increased skin tumorigenesis in mice lacking pi class glutathione S-transferases. *Proc Natl Acad Sci U S A* 1998;95(9):5275-80.
12. Mulder TP, Manni JJ, Roelofs HM, Peters WH, Wiersma A. Glutathione S-transferases and glutathione in human head and neck cancer. *Carcinogenesis* 1995;16(3):619-24.
13. Matthias C, Bockmuhl U, Jahnke V, Harries LW, Wolf CR, Jones PW, et al. The glutathione S-transferase GSTP1 polymorphism: effects on susceptibility to oral/pharyngeal and laryngeal carcinomas. *Pharmacogenetics* 1998;8(1):1-6.
14. Ha PK, Califano JA. The molecular biology of mucosal field cancerization of the head and neck. *Crit Rev Oral Biol Med* 2003;14(5):363-9.
15. Ogden GR, Chisholm DM, Morris AM, Stevenson JH. Overexpression of p53 in normal oral mucosa of oral cancer patients does not necessarily predict further malignant disease. *J Pathol* 1997;182(2):180-4.
16. Riethdorf S, Friedrich RE, Suhwold J, Ostwald C, Barten M, Gogacz P, et al. [p53 mutations and HPV infections in squamous epithelial carcinomas of the head-neck region. Long-term follow-up]. *Mund Kiefer Gesichtschir* 1998;2(1):30-4.
17. Riethdorf S, Friedrich RE, Ostwald C, Barten M, Gogacz P, Gundlach KK, et al. p53 gene mutations and HPV infection in primary head and neck squamous cell carcinomas do not correlate with overall survival: a long-term follow-up study. *J Oral Pathol Med* 1997;26(7):315-21.

18. Spano JP, Busson P, Atlan D, Bourhis J, Pignon JP, Esteban C, et al. Nasopharyngeal carcinomas: an update. *Eur J Cancer* 2003;39(15):2121-35.
19. Sunderman FW, Jr. Nasal toxicity, carcinogenicity, and olfactory uptake of metals. *Ann Clin Lab Sci* 2001;31(1):3-24.
20. Klintenberg C, Olofsson J, Hellquist H, Sokjer H. Adenocarcinoma of the ethmoid sinuses. A review of 28 cases with special reference to wood dust exposure. *Cancer* 1984;54(3):482-8.
21. Ward MH, Pan WH, Cheng YJ, Li FH, Brinton LA, Chen CJ, et al. Dietary exposure to nitrite and nitrosamines and risk of nasopharyngeal carcinoma in Taiwan. *Int J Cancer* 2000;86(5):603-9.
22. Foulkes WD, Brunet JS, Kowalski LP, Narod SA, Franco EL. Family history of cancer is a risk factor for squamous cell carcinoma of the head and neck in Brazil: a case-control study. *Int J Cancer* 1995;63(6):769-73.
23. Shao JY, Zeng WF, Zeng YX. [Molecular genetic progression on nasopharyngeal carcinoma]. *Ai Zheng* 2002;21(1):1-10.
24. Lei Z, Liu Q. [Bcl-2 oncoprotein expression in head and neck malignant neoplasm]. *Lin Chuang Er Bi Yan Hou Ke Za Zhi* 1998;12(4):171-3.
25. Gollin SM. Chromosomal alterations in squamous cell carcinomas of the head and neck: window to the biology of disease. *Head Neck* 2001;23(3):238-53.
26. Weber A, Wittekind C, Tannapfel A. Genetic and epigenetic alterations of 9p21 gene products in benign and malignant tumors of the head and neck. *Pathol Res Pract* 2003;199(6):391-7.
27. Wong TS, Man MW, Lam AK, Wei WI, Kwong YL, Yuen AP. The study of p16 and p15 gene methylation in head and neck squamous cell carcinoma and their quantitative evaluation in plasma by real-time PCR. *Eur J Cancer* 2003;39(13):1881-7.
28. Liggett WH, Jr., Sewell DA, Rocco J, Ahrendt SA, Koch W, Sidransky D. p16 and p16 beta are potent growth suppressors of head and neck squamous carcinoma cells in vitro. *Cancer Res* 1996;56(18):4119-23.
29. Chakraborty SB, Dasgupta S, Roy A, Sengupta A, Ray B, Roychoudhury S, et al. Differential deletions in 3p are associated with the development of head and neck squamous cell carcinoma in Indian patients. *Cancer Genet Cytogenet* 2003;146(2):130-8.
30. Virgilio L, Shuster M, Gollin SM, Veronese ML, Ohta M, Huebner K, et al. FHIT gene alterations in head and neck squamous cell carcinomas. *Proc Natl Acad Sci U S A* 1996;93(18):9770-5.
31. Gotte K, Riedel F, Neubauer J, Schafer C, Coy JF, Hormann K. The relationship between allelic imbalance on 17p, p53 mutation and p53 overexpression in head and neck cancer. *Int J Oncol* 2001;19(2):331-6.
32. Gupta VK, Schmidt AP, Pashia ME, Sunwoo JB, Scholnick SB. Multiple regions of deletion on chromosome arm 13q in head-and-neck squamous-cell carcinoma. *Int J Cancer* 1999;84(5):453-7.
33. Nawroz-Danish HM, Koch WM, Westra WH, Yoo G, Sidransky D. Lack of BRCA2 alterations in primary head and neck squamous cell carcinoma. *Otolaryngol Head Neck Surg* 1998;119(1):21-5.
34. Hamel N, Manning A, Black MJ, Tonin PN, Foulkes WD. An absence of founder BRCA2 mutations in individuals with squamous cell carcinoma of the head and neck. *Int J Cancer* 1999;83(6):803-4.
35. Muller D, Millon R, Velten M, Bronner G, Jung G, Engelmann A, et al. Amplification of 11q13 DNA markers in head and neck squamous cell carcinomas: correlation with clinical outcome. *Eur J Cancer* 1997;33(13):2203-10.

36. Meredith SD, Levine PA, Burns JA, Gaffey MJ, Boyd JC, Weiss LM, et al. Chromosome 11q13 amplification in head and neck squamous cell carcinoma. Association with poor prognosis. *Arch Otolaryngol Head Neck Surg* 1995;121(7):790-4.
37. Xu J, Gimenez-Conti IB, Cunningham JE, Collet AM, Luna MA, Lanfranchi HE, et al. Alterations of p53, cyclin D1, Rb, and H-ras in human oral carcinomas related to tobacco use. *Cancer* 1998;83(2):204-12.
38. Hoa M, Davis SL, Ames SJ, Spanjaard RA. Amplification of wild-type K-ras promotes growth of head and neck squamous cell carcinoma. *Cancer Res* 2002;62(24):7154-6.
39. Christensen ME. The EGF receptor system in head and neck carcinomas and normal tissues. Immunohistochemical and quantitative studies. *Dan Med Bull* 1998;45(2):121-34.
40. Solorzano CC, Jones SC, Pettitjean M, O'Daniel TG, Auffenberg T, Woost PG, et al. Inhibition of transforming growth factor alpha stimulation of human squamous cell carcinoma of the head and neck with anti-TGF-alpha antibodies and tyrphostin. *Ann Surg Oncol* 1997;4(8):670-84.
41. Bergler W, Petroianu G, Juncker C, Hormann K. Correlation of transforming growth factor alpha and epidermal growth factor receptor in oropharyngeal carcinomas. *Acta Otolaryngol* 1996;116(3):486-9.
42. Ninck S, Reisser C, Dyckhoff G, Helmke B, Bauer H, Herold-Mende C. Expression profiles of angiogenic growth factors in squamous cell carcinomas of the head and neck. *Int J Cancer* 2003;106(1):34-44.
43. Khan AJ, King BL, Smith BD, Smith GL, DiGiovanna MP, Carter D, et al. Characterization of the HER-2/neu oncogene by immunohistochemical and fluorescence in situ hybridization analysis in oral and oropharyngeal squamous cell carcinoma. *Clin Cancer Res* 2002;8(2):540-8.
44. Uno M, Otsuki T, Kurebayashi J, Sakaguchi H, Isozaki Y, Ueki A, et al. Anti-HER2-antibody enhances irradiation-induced growth inhibition in head and neck carcinoma. *Int J Cancer* 2001;94(4):474-9.
45. Mao L, El-Naggar AK, Fan YH, Lee JS, Lippman SM, Kayser S, et al. Telomerase activity in head and neck squamous cell carcinoma and adjacent tissues. *Cancer Res* 1996;56(24):5600-4.
46. Mutirangura A, Supiyaphun P, Trirekapan S, Sriuranpong V, Sakuntabhai A, Yenrudi S, et al. Telomerase activity in oral leukoplakia and head and neck squamous cell carcinoma. *Cancer Res* 1996;56(15):3530-3.
47. Califano J, Ahrendt SA, Meininger G, Westra WH, Koch WM, Sidransky D. Detection of telomerase activity in oral rinses from head and neck squamous cell carcinoma patients. *Cancer Res* 1996;56(24):5720-2.
48. Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. *Cell* 1990;61(5):759-67.
49. Ah-See KW, Cooke TG, Pickford IR, Soutar D, Balmain A. An allelotype of squamous carcinoma of the head and neck using microsatellite markers. *Cancer Res* 1994;54(7):1617-21.
50. el-Naggar AK, Hurr K, Batsakis JG, Luna MA, Goepfert H, Huff V. Sequential loss of heterozygosity at microsatellite motifs in preinvasive and invasive head and neck squamous carcinoma. *Cancer Res* 1995;55(12):2656-9.
51. El-Naggar AK, Hurr K, Huff V, Clayman GL, Luna MA, Batsakis JG. Microsatellite instability in preinvasive and invasive head and neck squamous carcinoma. *Am J Pathol* 1996;148(6):2067-72.

52. Franklin WA, Veve R, Hirsch FR, Helfrich BA, Bunn PA, Jr. Epidermal growth factor receptor family in lung cancer and premalignancy. *Semin Oncol* 2002;29(1 Suppl 4):3-14.
53. Califano J, Westra WH, Meininger G, Corio R, Koch WM, Sidransky D. Genetic progression and clonal relationship of recurrent premalignant head and neck lesions. *Clin Cancer Res* 2000;6(2):347-52.
54. Califano J, van der Riet P, Westra W, Nawroz H, Clayman G, Piantadosi S, et al. Genetic progression model for head and neck cancer: implications for field cancerization. *Cancer Res* 1996;56(11):2488-92.
55. Yoo GH, Washington J, Piechocki M, Ensley J, Shibuya T, Oda D, et al. Progression of head and neck cancer in an in vitro model. *Arch Otolaryngol Head Neck Surg* 2000;126(11):1313-8.
56. Ha PK, Califano JA, 3rd. The molecular biology of laryngeal cancer. *Otolaryngol Clin North Am* 2002;35(5):993-1012.
57. Cross DS, Platt JL, Juhn SK, Bach FH, Adams GL. Tumor infiltrating lymphocytes in squamous cell carcinoma of the head and neck: mechanisms of enhancement using prostaglandin synthetase inhibitors. *Adv Exp Med Biol* 1997;400B:1013-24.
58. de Bree R, Roos JC, Plaizier MA, Quak JJ, van Kamp GJ, den Hollander W, et al. Selection of monoclonal antibody E48 IgG or U36 IgG for adjuvant radioimmunotherapy in head and neck cancer patients. *Br J Cancer* 1997;75(7):1049-60.
59. Van Hal NL, Van Dongen GA, Rood-Knippels EM, Van Der Valk P, Snow GB, Brakenhoff RH. Monoclonal antibody U36, a suitable candidate for clinical immunotherapy of squamous-cell carcinoma, recognizes a CD44 isoform. *Int J Cancer* 1996;68(4):520-7.
60. Brakenhoff RH, van Gog FB, Looney JE, van Walsum M, Snow GB, van Dongen GA. Construction and characterization of the chimeric monoclonal antibody E48 for therapy of head and neck cancer. *Cancer Immunol Immunother* 1995;40(3):191-200.
61. de Bree R, Roos JC, Quak JJ, den Hollander W, Snow GB, van Dongen GA. Radioimmunoscintigraphy and biodistribution of technetium-99m-labeled monoclonal antibody U36 in patients with head and neck cancer. *Clin Cancer Res* 1995;1(6):591-8.
62. van Gog FB, Brakenhoff RH, Snow GB, van Dongen GA. Rapid elimination of mouse/human chimeric monoclonal antibodies in nude mice. *Cancer Immunol Immunother* 1997;44(2):103-11.
63. Cortesina G, De Stefani A, Majore L, Forni G, Galeazzi E. [Loco-regional treatment with low and high doses of interleukin-2 of head and neck squamous cell carcinoma recurrences]. *Acta Otorhinolaryngol Ital* 1994;14(1):3-9.
64. Lang S, Zeidler R, Pauli C, Andratschke M, Wollenberg B. [IL-2 gene therapy in ENT carcinomas]. *Laryngorhinootologie* 2001;80(4):191-5.
65. Zatterstrom UK, Brun E, Willen R, Kjellen E, Wennerberg J. Tumor angiogenesis and prognosis in squamous cell carcinoma of the head and neck. *Head Neck* 1995;17(4):312-8.
66. Guang-Wu H, Sunagawa M, Jie-En L, Shimada S, Gang Z, Tokeshi Y, et al. The relationship between microvessel density, the expression of vascular endothelial growth factor (VEGF), and the extension of nasopharyngeal carcinoma. *Laryngoscope* 2000;110(12):2066-9.
67. Wilson RF, Morse MA, Pei P, Renner RJ, Schuller DE, Robertson FM, et al. Endostatin inhibits migration and invasion of head and neck squamous cell carcinoma cells. *Anticancer Res* 2003;23(2B):1289-95.

68. Porkka K. [Endostatin--light on the treatment of cancer in mice]. *Duodecim* 1997;113(22):2241-3.
69. Rocco JW, Li D, Liggett WH, Jr., Duan L, Saunders JK, Jr., Sidransky D, et al. p16INK4A adenovirus-mediated gene therapy for human head and neck squamous cell cancer. *Clin Cancer Res* 1998;4(7):1697-704.
70. Goebel EA, Davidson BL, Graham SM, Kern JA. Tumor reduction in vivo after adenoviral mediated gene transfer of the herpes simplex virus thymidine kinase gene and ganciclovir treatment in human head and neck squamous cell carcinoma. *Otolaryngol Head Neck Surg* 1998;119(4):331-6.
71. Konopleva M, Zhao S, Hu W, Jiang S, Snell V, Weidner D, et al. The anti-apoptotic genes Bcl-X(L) and Bcl-2 are over-expressed and contribute to chemoresistance of non-proliferating leukaemic CD34+ cells. *Br J Haematol* 2002;118(2):521-34.
72. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. *Nature* 2001;409(6822):860-921.
73. Brown PO, Botstein D. Exploring the new world of the genome with DNA microarrays. *Nat Genet* 1999;21(1 Suppl):33-7.
74. Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. Light-directed, spatially addressable parallel chemical synthesis. *Science* 1991;251(4995):767-73.
75. Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SP. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc Natl Acad Sci U S A* 1994;91(11):5022-6.
76. Nuwaysir EF, Bittner M, Trent J, Barrett JC, Afshari CA. Microarrays and toxicology: the advent of toxicogenomics. *Mol Carcinog* 1999;24(3):153-9.
77. Lindblad-Toh K, Tanenbaum DM, Daly MJ, Winchester E, Lui WO, Villapakkam A, et al. Loss-of-heterozygosity analysis of small-cell lung carcinomas using single-nucleotide polymorphism arrays. *Nat Biotechnol* 2000;18(9):1001-5.
78. McClintick JN, Jerome RE, Nicholson CR, Crabb DW, Edenberg HJ. Reproducibility of oligonucleotide arrays using small samples. *BMC Genomics* 2003;4(1):4.
79. Okamoto T, Suzuki T, Yamamoto N. Microarray fabrication with covalent attachment of DNA using bubble jet technology. *Nat Biotechnol* 2000;18(4):438-41.
80. Hunter L, Taylor RC, Leach SM, Simon R. GEST: a gene expression search tool based on a novel Bayesian similarity metric. *Bioinformatics* 2001;17 Suppl 1:S115-22.
81. Quackenbush J. Computational analysis of microarray data. *Nat Rev Genet* 2001;2(6):418-27.
82. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 1998;95(25):14863-8.
83. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A* 1999;96(6):2907-12.
84. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. *Nat Genet* 1999;22(3):281-5.
85. DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, et al. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat Genet* 1996;14(4):457-60.
86. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon

tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci U S A* 1999;96(12):6745-50.
87. Alizadeh AA, Staudt LM. Genomic-scale gene expression profiling of normal and malignant immune cells. *Curr Opin Immunol* 2000;12(2):219-25.
88. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 2000;403(6769):503-11.
89. Sultan M, Wigle DA, Cumbaa CA, Maziarz M, Glasgow J, Jurisica I. Binary Tree-Structured Vector Quantization Approach to Clustering and Visualising Microarray Data. *Bioinformatics* 2002;1(1):1-9.
90. Sultan M, Wigle DA, Cumbaa CA, Maziarz M, Glasgow J, Tsao MS, et al. Binary tree-structured vector quantization approach to clustering and visualizing microarray data. *Bioinformatics* 2002;18 Suppl 1:S111-S119.
91. Gersho A GR. Vector Quantization and signal compression. *Kluwer* 1992.
92. Vesanto. SOM-Based Data Visualisation Methods. *Intelligence Data Analyis* 1999;3:111-129.
93. Tibshirani R, Hastie T, Eisen MB, Ross D, Botstein D, Brown PO. Clustering methods for the analysis of DNA microarray data. *Technical Report,Department of Statistics, Stanford University* 1999.
94. Wigle DA, Jurisica I, Radulovich N, Pintilie M, Rossant J, Liu N, et al. Molecular profiling of non-small cell lung cancer and correlation with disease-free survival. *Cancer Res* 2002;62(11):3005-8.
95. Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, et al. Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet* 2000;24(3):227-35.
96. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999;286(5439):531-7.
97. Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, et al. Gene-expression profiles in hereditary breast cancer. *N Engl J Med* 2001;344(8):539-48.
98. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A* 2001;98(19):10869-74.
99. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. *Nature* 2000;406(6797):747-52.
100. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 2000;403(6769):503-11.
101. Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, et al. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* 2000;406(6795):536-40.
102. Khan J, Simon R, Bittner M, Chen Y, Leighton SB, Pohida T, et al. Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res* 1998;58(22):5009-13.
103. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci U S A* 2001;98(24):13790-5.
104. Garber ME, Troyanskaya OG, Schluens K, Petersen S, Thaesler Z, Pacyna-Gengelbach M, et al. Diversity of gene expression in adenocarcinoma of the lung. *Proc Natl Acad Sci U S A* 2001;98(24):13784-9.

105. Dhanasekaran SM, Barrette TR, Ghosh D, Shah R, Varambally S, Kurachi K, et al. Delineation of prognostic biomarkers in prostate cancer. *Nature* 2001;412(6849):822-6.
106. Hacia JG. Resequencing and mutational analysis using oligonucleotide microarrays. *Nat Genet* 1999;21(1 Suppl):42-7.
107. Debouck C, Goodfellow PN. DNA microarrays in drug discovery and development. *Nat Genet* 1999;21(1 Suppl):48-50.
108. Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, et al. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat Genet* 1999;23(1):41-6.
109. Beheshti B, Park PC, Braude I, Squire JA. Microarray CGH. *Methods Mol Biol* 2002;204:191-207.
110. Mantripragada KK, Buckley PG, Benetkiewicz M, De Bustos C, Hirvela C, Jarbo C, et al. High-resolution profiling of an 11 Mb segment of human chromosome 22 in sporadic schwannoma using array-CGH. *Int J Oncol* 2003;22(3):615-22.
111. Beheshti B, Braude I, Marrano P, Thorner P, Zielenska M, Squire JA. Chromosomal localization of DNA amplifications in neuroblastoma tumors using cDNA microarray comparative genomic hybridization. *Neoplasia* 2003;5(1):53-62.
112. Kraus J, Pantel K, Pinkel D, Albertson DG, Speicher MR. High-resolution genomic profiling of occult micrometastatic tumor cells. *Genes Chromosomes Cancer* 2003;36(2):159-66.
113. Belbin TJ, Singh B, Barber I, Socci N, Wenig B, Smith R, et al. Molecular classification of head and neck squamous cell carcinoma using cDNA microarrays. *Cancer Res* 2002;62(4):1184-90.
114. Villaret DB, Wang T, Dillon D, Xu J, Sivam D, Cheever MA, et al. Identification of genes overexpressed in head and neck squamous cell carcinoma using a combination of complementary DNA subtraction and microarray analysis. *Laryngoscope* 2000;110(3 Pt 1):374-81.
115. Alevizos I, Mahadevappa M, Zhang X, Ohyama H, Kohno Y, Posner M, et al. Oral cancer in vivo gene expression profiling assisted by laser capture microdissection and microarray analysis. *Oncogene* 2001;20(43):6196-204.
116. Leethanakul C, Patel V, Gillespie J, Pallente M, Ensley JF, Koontongkaew S, et al. Distinct pattern of expression of differentiation and growth-related genes in squamous cell carcinomas of the head and neck revealed by the use of laser capture microdissection and cDNA arrays. *Oncogene* 2000;19(28):3220-4.
117. Squire JA, Bayani J, Luk C, Unwin L, Tokunaga J, MacMillan C, et al. Molecular cytogenetic analysis of head and neck squamous cell carcinoma: By comparative genomic hybridization, spectral karyotyping, and expression array analysis. *Head Neck* 2002;24(9):874-87.
118. Casiglia J, Woo SB. A comprehensive review of oral cancer. *Gen Dent* 2001;49(1):72-82.
119. Balaram P, Sridhar H, Rajkumar T, Vaccarella S, Herrero R, Nandakumar A, et al. Oral cancer in southern India: the influence of smoking, drinking, paan-chewing and oral hygiene. *Int J Cancer* 2002;98(3):440-5.
120. Fossion E, De Coster D, Ehlinger P. [Oral cancer: epidemiology and prognosis]. *Rev Belge Med Dent* 1994;49(1):9-22.
121. Maier H, Dietz A, Gewelke U, Heller WD, Weidauer H. Tobacco and alcohol and the risk of head and neck cancer. *Clin Investig* 1992;70(3-4):320-7.
122. Landis SH, Murray T, Bolden S, Wingo PA. Cancer statistics, 1999. *CA Cancer J Clin* 1999;49(1):8-31, 1.

123. Houck JR, Medina JE. Management of cervical lymph nodes in squamous carcinomas of the head and neck. *Semin Surg Oncol* 1995;11(3):228-39.
124. Tankere F, Camproux A, Barry B, Guedon C, Depondt J, Gehanno P. Prognostic value of lymph node involvement in oral cancers: a study of 137 cases. *Laryngoscope* 2000;110(12):2061-5.
125. Smith BD, Haffty BG. Molecular markers as prognostic factors for local recurrence and radioresistance in head and neck squamous cell carcinoma. *Radiat Oncol Investig* 1999;7(3):125-44.
126. Scully C, Field JK, Tanzawa H. Genetic aberrations in oral or head and neck squamous cell carcinoma 3: clinico-pathological applications. *Oral Oncol* 2000;36(5):404-13.
127. Shah JP. Patterns of cervical lymph node metastasis from squamous carcinomas of the upper aerodigestive tract. *Am J Surg* 1990;160(4):405-9.
128. Kramer D, Durham JS, Jackson S, Brookes J. Management of the neck in N0 squamous cell carcinoma of the oral cavity. *J Otolaryngol* 2001;30(5):283-8.
129. Huang Y, Prasad M, Lemon WJ, Hampel H, Wright FA, Kornacker K, et al. Gene expression in papillary thyroid carcinoma reveals highly consistent profiles. *Proc Natl Acad Sci U S A* 2001;98(26):15044-9.
130. Homma A, Furuta Y, Oridate N, Nakano Y, Kohashi G, Yagi K, et al. Prognostic significance of clinical parameters and biological markers in patients with squamous cell carcinoma of the head and neck treated with concurrent chemoradiotherapy. *Clin Cancer Res* 1999;5(4):801-6.
131. Al Moustafa AE, Alaoui-Jamali MA, Batist G, Hernandez-Perez M, Serruya C, Alpert L, et al. Identification of genes associated with head and neck carcinogenesis by cDNA microarray comparison between matched primary normal epithelial and squamous carcinoma cells. *Oncogene* 2002;21(17):2634-40.
132. Huang Q, Yu GP, McCormick SA, Mo J, Datta B, Mahimkar M, et al. Genetic differences detected by comparative genomic hybridization in head and neck squamous cell carcinomas from different tumor sites: construction of oncogenetic trees for tumor progression. *Genes Chromosomes Cancer* 2002;34(2):224-33.
133. Wang E, Miller LD, Ohnmacht GA, Liu ET, Marincola FM. High-fidelity mRNA amplification for gene profiling. *Nat Biotechnol* 2000;18(4):457-9.
134. Baugh LR, Hill AA, Brown EL, Hunter CP. Quantitative analysis of mRNA amplification by in vitro transcription. *Nucleic Acids Res* 2001;29(5):E29.
135. Iscove NN, Barbara M, Gu M, Gibson M, Modi C, Winegarden N. Representation is faithfully preserved in global cDNA amplified exponentially from sub-picogram quantities of mRNA. *Nat Biotechnol* 2002;20(9):940-3.
136. Hu L, Wang J, Baggerly K, Wang H, Fuller GN, Hamilton SR, et al. Obtaining reliable information from minute amounts of RNA using cDNA microarrays. *BMC Genomics* 2002;3(1):16.
137. Makrigiorgos GM, Chakrabarti S, Zhang Y, Kaur M, Price BD. A PCR-based amplification method retaining the quantitative difference between two complex genomes. *Nat Biotechnol* 2002;20(9):936-9.
138. Chomczynski P, Sacchi N. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal Biochem* 1987;162(1):156-9.
139. van Houten VM, Snijders PJ, van den Brekel MW, Kummer JA, Meijer CJ, van Leeuwen B, et al. Biological evidence that human papillomaviruses are etiologically involved in a subgroup of head and neck squamous cell carcinomas. *Int J Cancer* 2001;93(2):232-5.

140. Lee S, Baek M, Yang H, Bang YJ, Kim WH, Ha JH, et al. Identification of genes differentially expressed between gastric cancers and normal gastric mucosa with cDNA microarrays. *Cancer Lett* 2002;184(2):197-206.
141. Chiesa F, Mauri S, Tradati N, Calabrese L, Giugliano G, Ansarin M, et al. Surfing prognostic factors in head and neck cancer at the millennium. *Oral Oncol* 1999;35(6):590-6.
142. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002;415(6871):530-6.
143. Ben-Dor A, Shamir R, Yakhini Z. Clustering gene expression patterns. *J Comput Biol* 1999;6(3-4):281-97.
144. Maniotis AJ, Folberg R, Hess A, Seftor EA, Gardner LM, Pe'er J, et al. Vascular channel formation by human melanoma cells in vivo and in vitro: vasculogenic mimicry. *Am J Pathol* 1999;155(3):739-52.
145. Weyers W, Euler M, Diaz-Cascajo C, Schill WB, Bonczkowitz M. Classification of cutaneous malignant melanoma: a reassessment of histopathologic criteria for the distinction of different types. *Cancer* 1999;86(2):288-99.
146. Evangelou A, Letarte M, I. J, Sultan M, Murphy KJ, Rosen B, et al. Loss of coordinated androgen regulation in nonmalignant ovarian epithelial cells with BRCA1/2 mutations and ovarian cancer cells. *Cancer Research* 2003;63(in press).
147. Kohonen T, Somervuo P. How to make large self-organizing maps for nonvectorial data. *Neural Netw* 2002;15(8-9):945-52.
148. Kohonen T. Comparison of SOM point densities based on different criteria. *Neural Comput* 1999;11(8):2081-95.
149. Patane G, Russo M. The enhanced LBG algorithm. *Neural Netw* 2001;14(9):1219-37.
150. Poetsch M, Kleist B, Lorenz G, Herrmann FH. Different numerical chromosomal aberrations detected by FISH in oropharyngeal, hypopharyngeal and laryngeal squamous cell carcinoma. *Histopathology* 1999;34(3):234-40.
151. Chambers AF, Groom AC, MacDonald IC. Dissemination and growth of cancer cells in metastatic sites. *Nat Rev Cancer* 2002;2(8):563-72.
152. Steinhart H, Bohlender J, Iro H, Jung V, Constantinidis J, Gebhart E, et al. DNA amplification on chromosome 7q in squamous cell carcinoma of the tongue. *Int J Oncol* 2001;19(4):851-5.
153. Freier K, Joos S, Flechtenmacher C, Devens F, Benner A, Bosch FX, et al. Tissue microarray analysis reveals site-specific prevalence of oncogene amplifications in head and neck squamous cell carcinoma. *Cancer Res* 2003;63(6):1179-82.
154. Baatenburg de Jong RJ, Hermans J, Molenaar J, Briaire JJ, le Cessie S. Prediction of survival in patients with head and neck cancer. *Head Neck* 2001;23(9):718-24.
155. Simpson DA, Feeney S, Boyle C, Stitt AW. Retinal VEGF mRNA measured by SYBR green I fluorescence: A versatile approach to quantitative PCR. *Mol Vis* 2000;6:178-83.
156. Thellin O, Zorzi W, Lakaye B, De Borman B, Coumans B, Hennen G, et al. Housekeeping genes as internal standards: use and limits. *J Biotechnol* 1999;75(2-3):291-5.
157. Bustin SA. Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. *J Mol Endocrinol* 2000;25(2):169-93.
158. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* 2001;25(4):402-8.

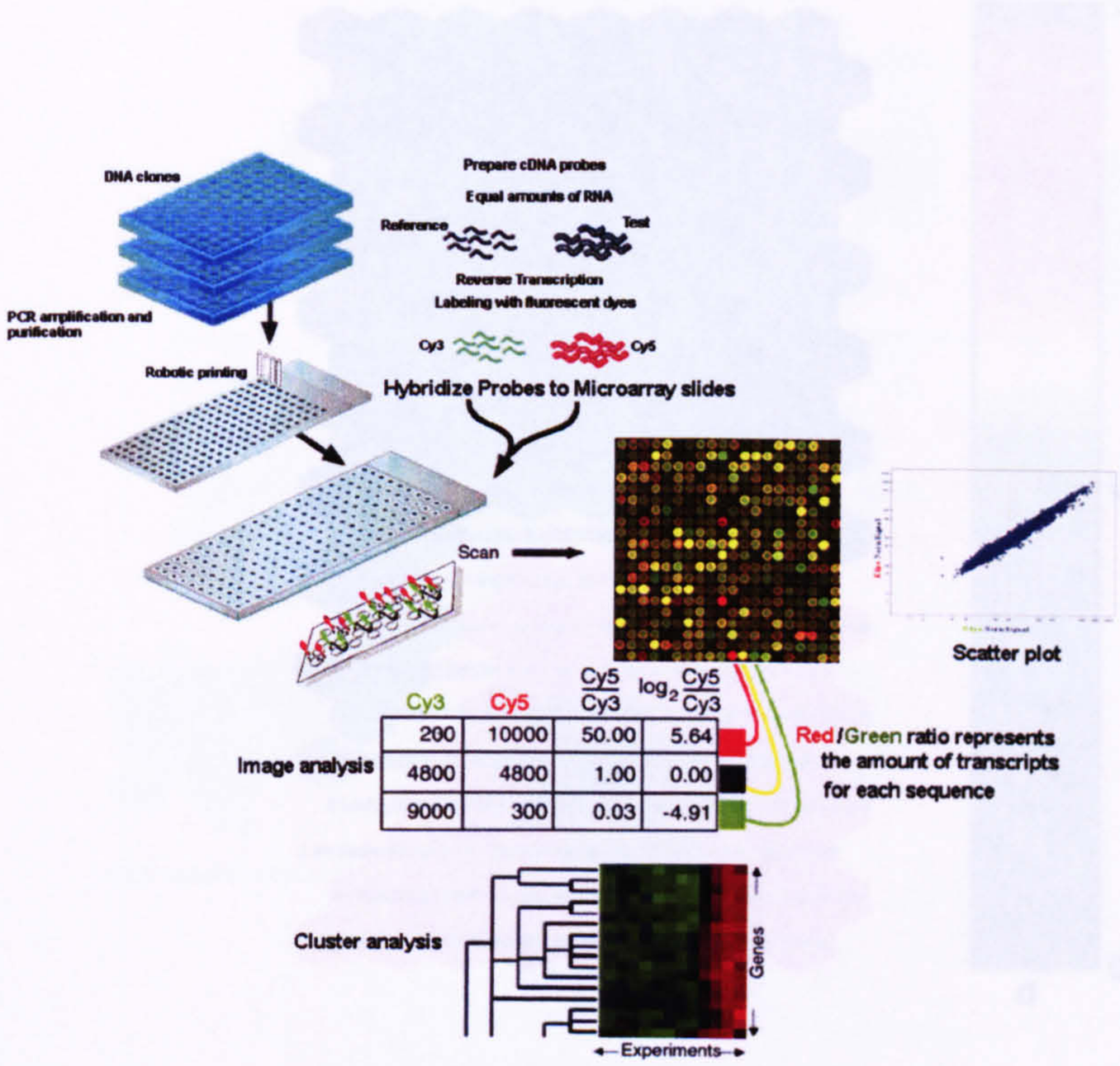
159. Klijanienko J, el-Naggar A, De Braud F, Micheau C, Janot F, Luboinski B, et al. Keratins 6, 13 and 19. Differential expression in squamous cell carcinoma of the head and neck. *Anal Quant Cytol Histol* 1993;15(5):335-40.
160. Leethanakul C, Patel V, Gillespie J, Shillitoe E, Kellman RM, Ensley JF, et al. Gene expression profiles in squamous cell carcinomas of the oral cavity: use of laser capture microdissection for the construction and analysis of stage-specific cDNA libraries. *Oral Oncol* 2000;36(5):474-83.
161. Opdenakker G, Fiten P, Nys G, Froyen G, Van Roy N, Speleman F, et al. The human MCP-3 gene (SCYA7): cloning, sequence analysis, and assignment to the C-C chemokine gene cluster on chromosome 17q11.2-q12. *Genomics* 1994;21(2):403-8.
162. Leethanakul C, Knezevic V, Patel V, Amornphimoltham P, Gillespie J, Shillitoe EJ, et al. Gene discovery in oral squamous cell carcinoma through the Head and Neck Cancer Genome Anatomy Project: confirmation by microarray analysis. *Oral Oncol* 2003;39(3):248-58.
163. Heiskala M, Peterson PA, Yang Y. The roles of claudin superfamily proteins in paracellular transport. *Traffic* 2001;2(2):93-8.
164. Hwang D, Alevizos I, Schmitt WA, Misra J, Ohyama H, Todd R, et al. Genomic dissection for characterization of cancerous oral epithelium tissues using transcription profiling. *Oral Oncol* 2003;39(3):259-68.
165. Inohara N, Koseki T, Chen S, Wu X, Nunez G. CIDE, a novel family of cell death activators with homology to the 45 kDa subunit of the DNA fragmentation factor. *Embo J* 1998;17(9):2526-33.
166. Chen Z, Guo K, Toh SY, Zhou Z, Li P. Mitochondria localization and dimerization are required for CIDE-B to induce apoptosis. *J Biol Chem* 2000;275(30):22619-22.
167. Francioso F, Carinci F, Tosi L, Scapoli L, Pezzetti F, Passerella E, et al. Identification of differentially expressed genes in human salivary gland tumors by DNA microarrays. *Mol Cancer Ther* 2002;1(7):533-8.
168. Bennett EP, Hassan H, Hollingsworth MA, Clausen H. A novel human UDP-N-acetyl-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase, GalNAc-T7, with specificity for partial GalNAc-glycosylated acceptor substrates. *FEBS Lett* 1999;460(2):226-30.
169. Sutherlin ME, Nishimori I, Caffrey T, Bennett EP, Hassan H, Mandel U, et al. Expression of three UDP-N-acetyl-alpha-D-galactosamine:polypeptide GalNAc N-acetylgalactosaminyltransferases in adenocarcinoma cell lines. *Cancer Res* 1997;57(21):4744-8.
170. Ring HZ, Vameghi-Meyers V, Wang W, Crabtree GR, Francke U. Five SWI/SNF-related, matrix-associated, actin-dependent regulator of chromatin (SMARC) genes are dispersed in the human genome. *Genomics* 1998;51(1):140-3.
171. Fink TM, Vaesen M, Kratzin HD, Lichter P, Zimmer M. Localization of the gene encoding the putative human HLA class II associated protein (PHAPI) to chromosome 15q22.3-q23 by fluorescence in situ hybridization. *Genomics* 1995;29(1):309-10.
172. von Lindern M, van Baal S, Wiegant J, Raap A, Hagemeijer A, Grosveld G. Can, a putative oncogene associated with myeloid leukemogenesis, may be activated by fusion of its 3' half to different genes: characterization of the set gene. *Mol Cell Biol* 1992;12(8):3346-55.
173. Tasken K, Naylor SL, Solberg R, Jahnsen T. Mapping of the gene encoding the regulatory subunit RII alpha of cAMP-dependent protein kinase (locus PRKAR2A) to human chromosome region 3p21.3-p21.2. *Genomics* 1998;50(3):378-81.

174. Lozano Y, Taitz A, Petruzzelli GJ, Djordjevic A, Young MR. Prostaglandin E2-protein kinase A signaling and protein phosphatases-1 and -2A regulate human head and neck squamous cell carcinoma motility, adherence, and cytoskeletal organization. *Prostaglandins* 1996;51(1):35-48.
175. Folpe AL, Billings SD, McKenney JK, Walsh SV, Nusrat A, Weiss SW. Expression of claudin-1, a recently described tight junction-associated protein, distinguishes soft tissue perineurioma from potential mimics. *Am J Surg Pathol* 2002;26(12):1620-6.
176. Miwa N, Furuse M, Tsukita S, Niikawa N, Nakamura Y, Furukawa Y. Involvement of claudin-1 in the beta-catenin/Tcf signaling pathway and its frequent upregulation in human colorectal cancers. *Oncol Res* 2000;12(11-12):469-76.
177. Kramer F, White K, Kubbies M, Swisshelm K, Weber BH. Genomic organization of claudin-1 and its assessment in hereditary and sporadic breast cancer. *Hum Genet* 2000;107(3):249-56.
178. Schulz C, Petrig V, Wolf K, Kratzel K, Kohler M, Becker B, et al. Upregulation of MCAM in primary bronchial epithelial cells from patients with COPD. *Eur Respir J* 2003;22(3):450-6.
179. Evangelou A, Letarte M, Jurisica I, Sultan M, Murphy KJ, Rosen B, et al. Loss of coordinated androgen regulation in nonmalignant ovarian epithelial cells with BRCA1/2 mutations and ovarian cancer cells. *Cancer Res* 2003;63(10):2416-24.
180. Jurisica I, Wigle DA. Understanding biology through intelligent systems. *Genome Biol* 2002;3(11):reports4036.
181. Acton BM, Jurisicova A, Jurisica I, Casper RF. Alterations in mitochondrial membrane potential during preimplantation stages of mouse and human embryo development. *Mol Hum Reprod* 2004;10(1):23-32.
182. Bergamo NA, Rogatto SR, Poli-Frederico RC, Reis PP, Kowalski LP, Zielenska M, et al. Comparative genomic hybridization analysis detects frequent over-representation of DNA sequences at 3q, 7p, and 8q in head and neck carcinomas. *Cancer Genet Cytogenet* 2000;119(1):48-55.
183. Bockmuhl U, Schluns K, Schmidt S, Matthias S, Petersen I. Chromosomal alterations during metastasis formation of head and neck squamous cell carcinoma. *Genes Chromosomes Cancer* 2002;33(1):29-35.
184. Dasgupta S, Mukherjee N, Roy S, Roy A, Sengupta A, Roychowdhury S, et al. Mapping of the candidate tumor suppressor genes' loci on human chromosome 3 in head and neck squamous cell carcinoma of an Indian patient population. *Oral Oncol* 2002;38(1):6-15.
185. Forus A, Larramendy ML, Meza-Zepeda LA, Bjerkehagen B, Godager LH, Dahlberg AB, et al. Dedifferentiation of a well-differentiated liposarcoma to a highly malignant metastatic osteosarcoma: amplification of 12q14 at all stages and gain of 1q22-q24 associated with metastases. *Cancer Genet Cytogenet* 2001;125(2):100-11.
186. Rao PH, Murty VV, Louie DC, Chaganti RS. Nonsyntenic amplification of MYC with CDK4 and MDM2 in a malignant mixed tumor of salivary gland. *Cancer Genet Cytogenet* 1998;105(2):160-3.
187. El-Rifai W, Rutherford S, Knuutila S, Frierson HF, Jr., Moskaluk CA. Novel DNA copy number losses in chromosome 12q12--q13 in adenoid cystic carcinoma. *Neoplasia* 2001;3(3):173-8.
188. Campana WM, O'Brien JS, Hiraiwa M, Patton S. Secretion of prosaposin, a multifunctional protein, by breast cancer cells. *Biochim Biophys Acta* 1999;1427(3):392-400.
189. Misasi R, Sorice M, Di Marzio L, Campana WM, Molinari S, Cifone MG, et al. Prosaposin treatment induces PC12 entry in the S phase of the cell cycle and

prevents apoptosis: activation of ERKs and sphingosine kinase. *Faseb J* 2001;15(2):467-74.
190. Posner MR, Cavacini LA, Upton MP, Tillman KC, Gornstein ER, Norris CM, Jr. Surface membrane-expressed CD40 is present on tumor cells from squamous cell cancer of the head and neck in vitro and in vivo and regulates cell growth in tumor cell lines. *Clin Cancer Res* 1999;5(8):2261-70.
191. Loro LL, Ohlsson M, Vintermyr OK, Liavaag PG, Jonsson R, Johannessen AC. Maintained CD40 and loss of polarised CD40 ligand expression in oral squamous cell carcinoma. *Anticancer Res* 2001;21(1A):113-7.
192. Sabel MS, Yamada M, Kawaguchi Y, Chen FA, Takita H, Bankert RB. CD40 expression on human lung cancer correlates with metastatic spread. *Cancer Immunol Immunother* 2000;49(2):101-8.
193. Vonderheide RH, Dutcher JP, Anderson JE, Eckhardt SG, Stephans KF, Razvillas B, et al. Phase I study of recombinant human CD40 ligand in cancer patients. *J Clin Oncol* 2001;19(13):3280-7.
194. Nor JE, Christensen J, Liu J, Peters M, Mooney DJ, Strieter RM, et al. Up-Regulation of Bcl-2 in microvascular endothelial cells enhances intratumoral angiogenesis and accelerates tumor growth. *Cancer Res* 2001;61(5):2183-8.
195. Okamoto M, Reddy JK, Oyasu R. Tumorigenic conversion of a non-tumorigenic rat urothelial cell line by overexpression of H2O2-generating peroxisomal fatty acyl-CoA oxidase. *Int J Cancer* 1997;70(6):716-21.
196. Sundarrajan M, Fernandis AZ, Subrahmanyam G, Prabhudesai S, Krishnamurthy SC, Rao KV. Enhanced sequential expression of G1/S cyclins during experimental epatocarcinogenesis and tyrosine phosphorylation. *J Environ Pathol Toxicol Oncol* 2001;20(3):189-97.
197. Namazie A, Alavi S, Olopade OI, Pauletti G, Aghamohammadi N, Aghamohammadi M, et al. Cyclin D1 amplification and p16(MTS1/CDK4I) deletion correlate with poor prognosis in head and neck tumors. *Laryngoscope* 2002;112(3):472-81.
198. Nagy B, Tiszlavicz L, Eller J, Molnar J, Thurzo L. Ki-67, cyclin D1, p53 and bcl-2 expression in advanced head and neck cancer. *In Vivo* 2003;17(1):93-6.
199. Osman I, Sherman E, Singh B, Venkatraman E, Zelefsky M, Bosl G, et al. Alteration of p53 pathway in squamous cell carcinoma of the head and neck: impact on treatment outcome in patients treated with larynx preservation intent. *J Clin Oncol* 2002;20(13):2980-7.
200. P Oc, Rhys-Evans PH, Archer DJ, Eccles SA. C-erbB receptors in squamous cell carcinomas of the head and neck: clinical significance and correlation with matrix metalloproteinases and vascular endothelial growth factors. *Oral Oncol* 2002;38(1):73-80.
201. Doweck I, Barak M, Uri N, Greenberg E. The prognostic value of the tumour marker Cyfra 21-1 in carcinoma of head and neck and its role in early detection of recurrent disease. *Br J Cancer* 2000;83(12):1696-701.
202. Nylander K, Dabelsteen E, Hall PA. The p53 molecule and its prognostic role in squamous cell carcinomas of the head and neck. *J Oral Pathol Med* 2000;29(9):413-25.
203. Gleich LL, Li YQ, Wang X, Stambrook PJ, Gluckman JL. Variable genetic alterations and survival in head and neck cancer. *Arch Otolaryngol Head Neck Surg* 1999;125(9):949-52.

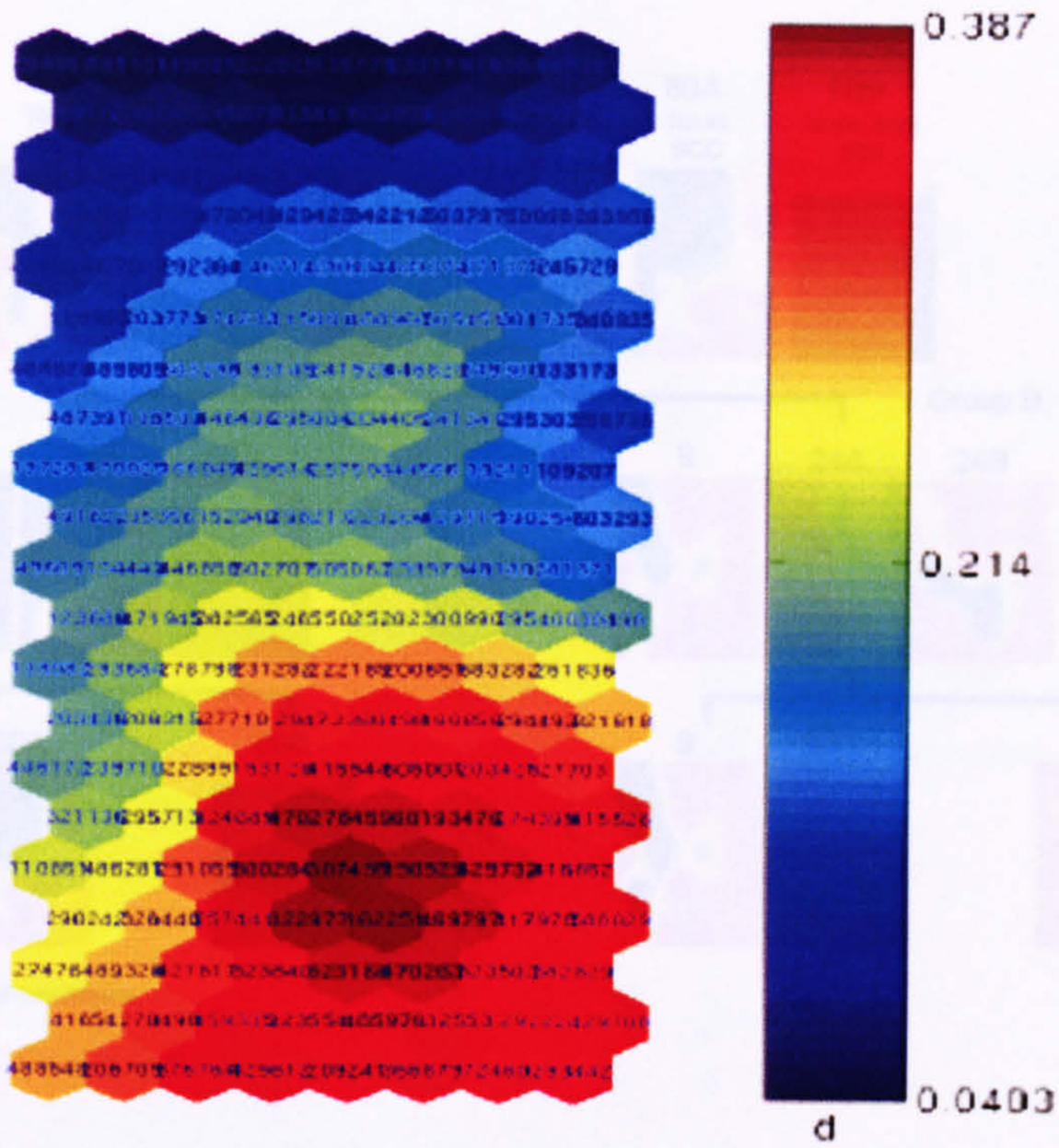
FIGURES

Figure 1



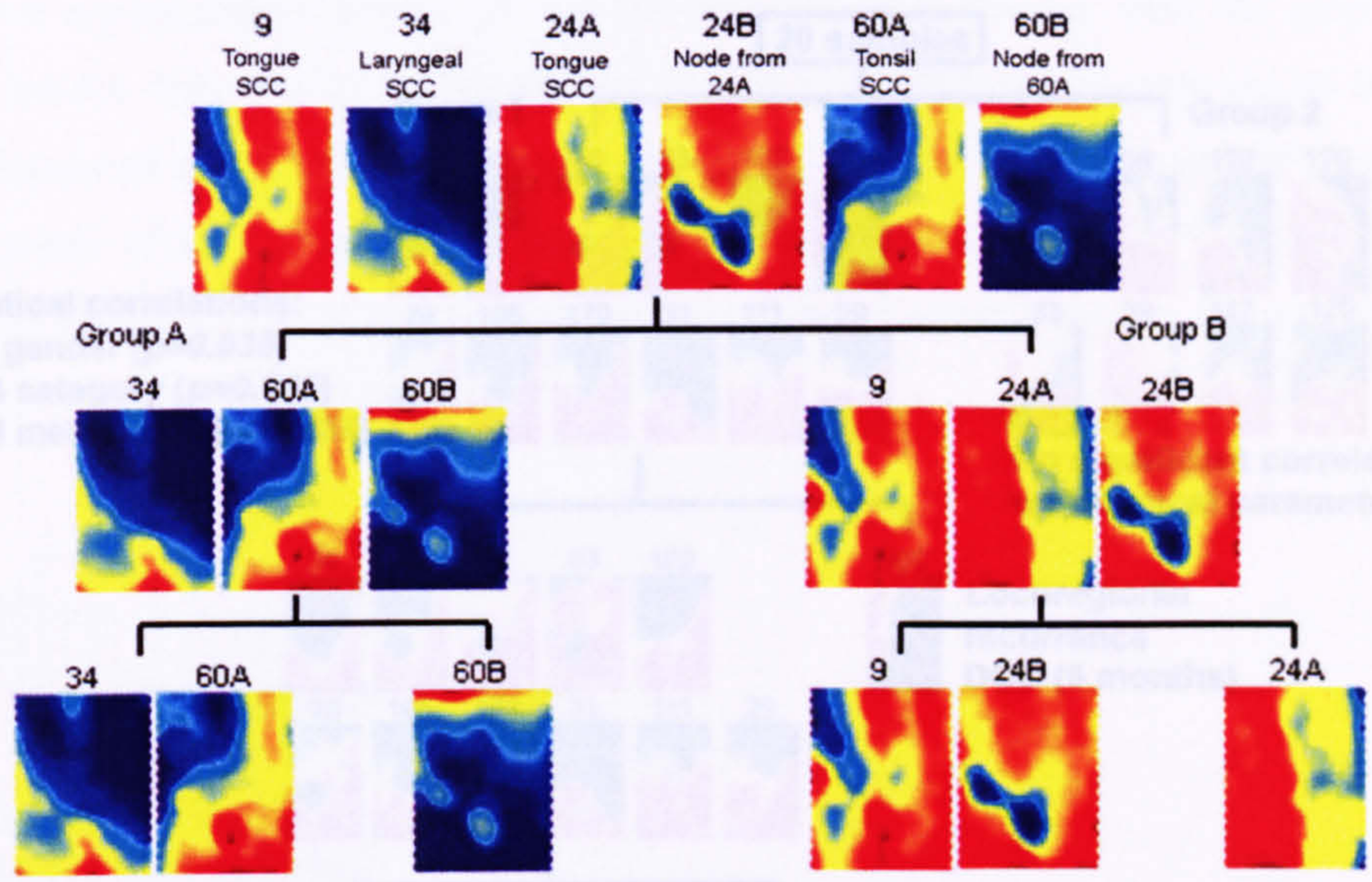
Schematic of the microarray technology and process

Figure 2:



A schematic of a single SOM component plane representing the gene expression profile of a single sample. Component planes are the planes of Voronoi Tessellations. Each hexagonal unit represents a gene (gene ID is seen within each heaxagon) to which a color is assigned. Red represents up-regulation, blue represents down-regulation, and yellow represents no change in expression. Genes with similar expression are clustered together.

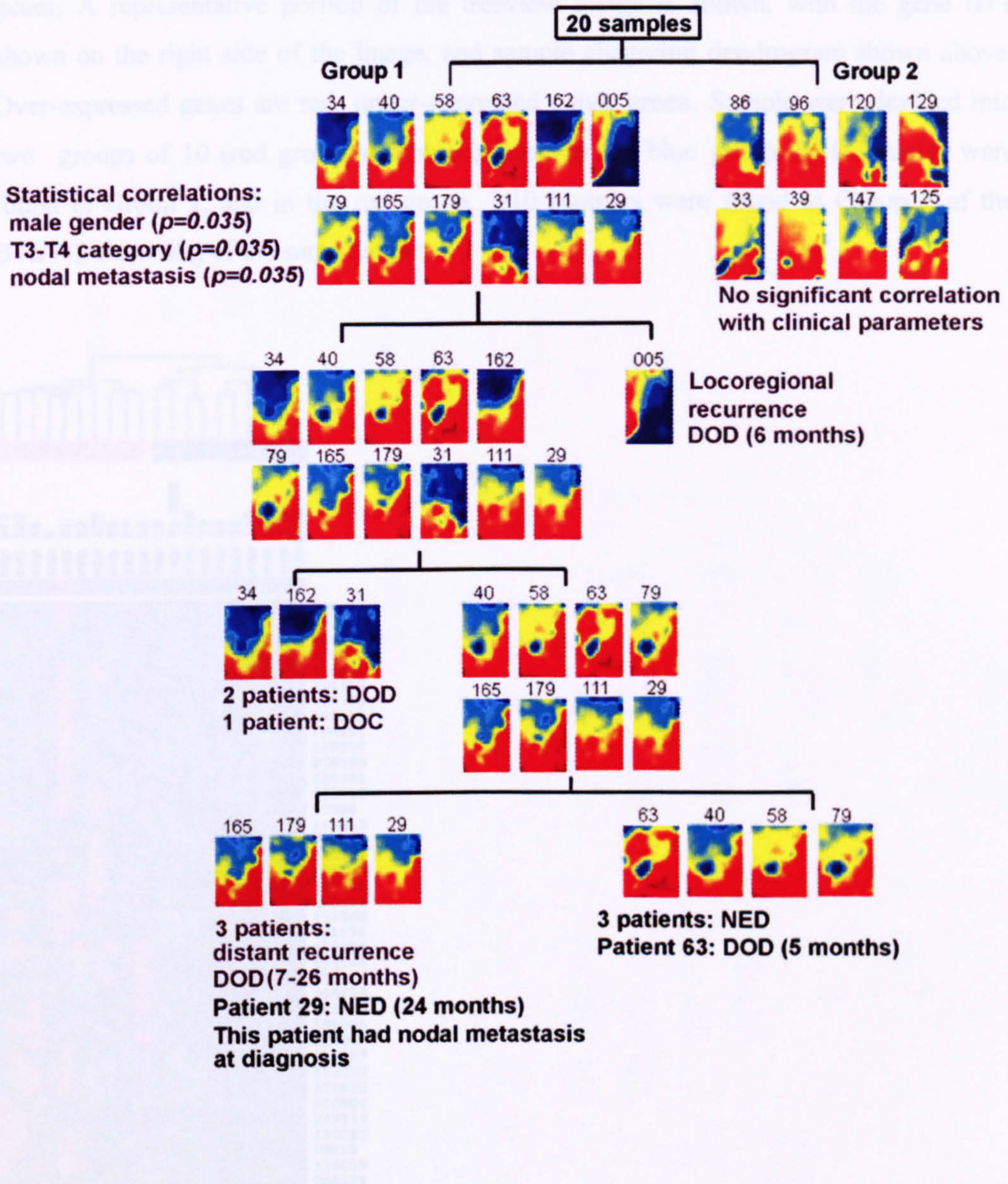
Figure 3:



SCC: squamous cell carcinoma

BTSVQ clustering of gene expression profiles from 6 HNSCC cell lines on the basis of 19,200 genes. A Self Organising Map (SOM) represents each sample. Comparison of expression profile between tumors is easily visualized by observing the color patterns. Samples with similar expression profiles are clustered at the nodes of the binary tree. Group A and B represent clusters containing primary and metastasis from the same patient. Group B contains all tongue carcinomas.

Figure 4



BTSVQ clustering of expression profiles from 20 oral carcinomas based on 19,200 genes. Each Self Organizing Map (SOM) represents the expression profile of a tumor. Composite profile represents clusters of genes that are similarly under- or overexpressed. Group 1 and 2 correlated with advanced stage tumors. DOD, died of disease; NED, no evidence of disease at last date of follow-up.

Figure 5

Unsupervised hierarchical clustering of 20 OSCC based on the expression of 2037 genes. A representative portion of the treeview image is shown, with the gene ID's shown on the right side of the image, and sample clustering dendrogram shown above. Over-expressed genes are red, under-expressed genes green. Sample were devided into two groups of 10 (red group and blue group). In the blue group, 9/10 samples were found in Group 1, and in the red group, 7/10 samples were found in Group 2 of the BTSVQ clustering of the same dataset.

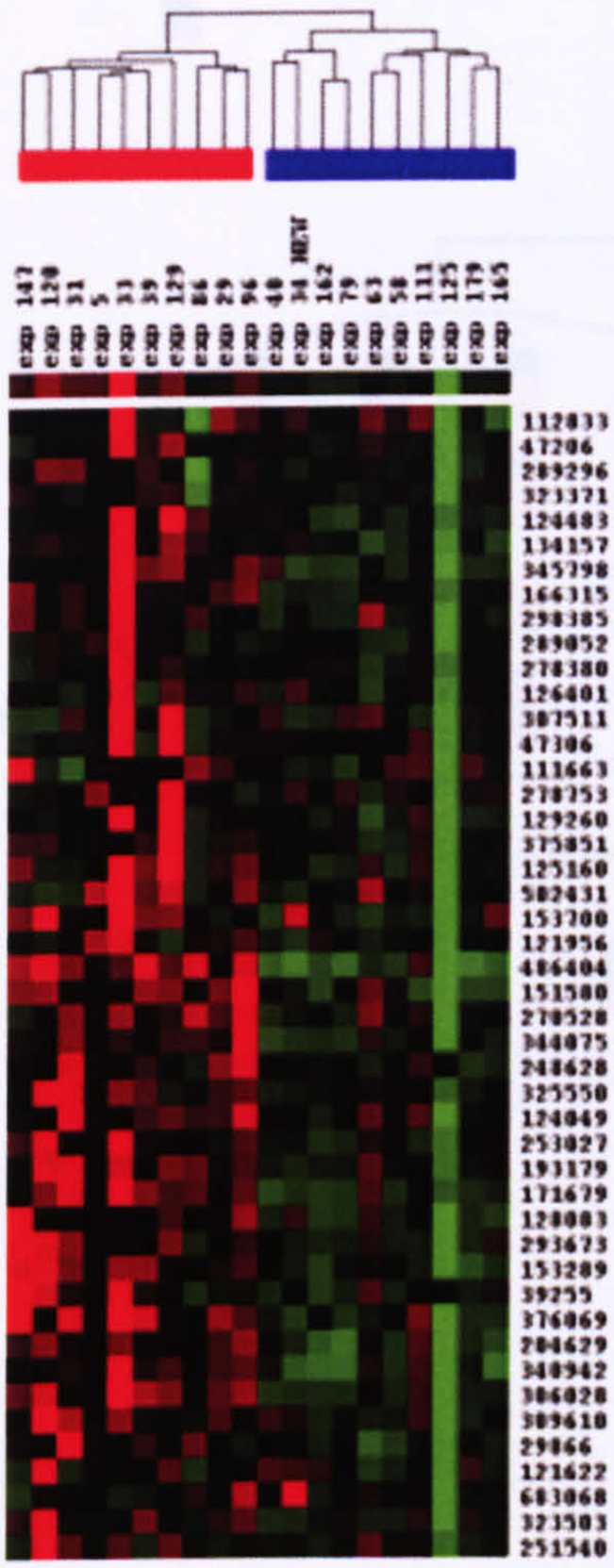


Figure 6

Clustering of 34 samples by BTSVQ. Group α & β correlated with nodal status and disease free survival.

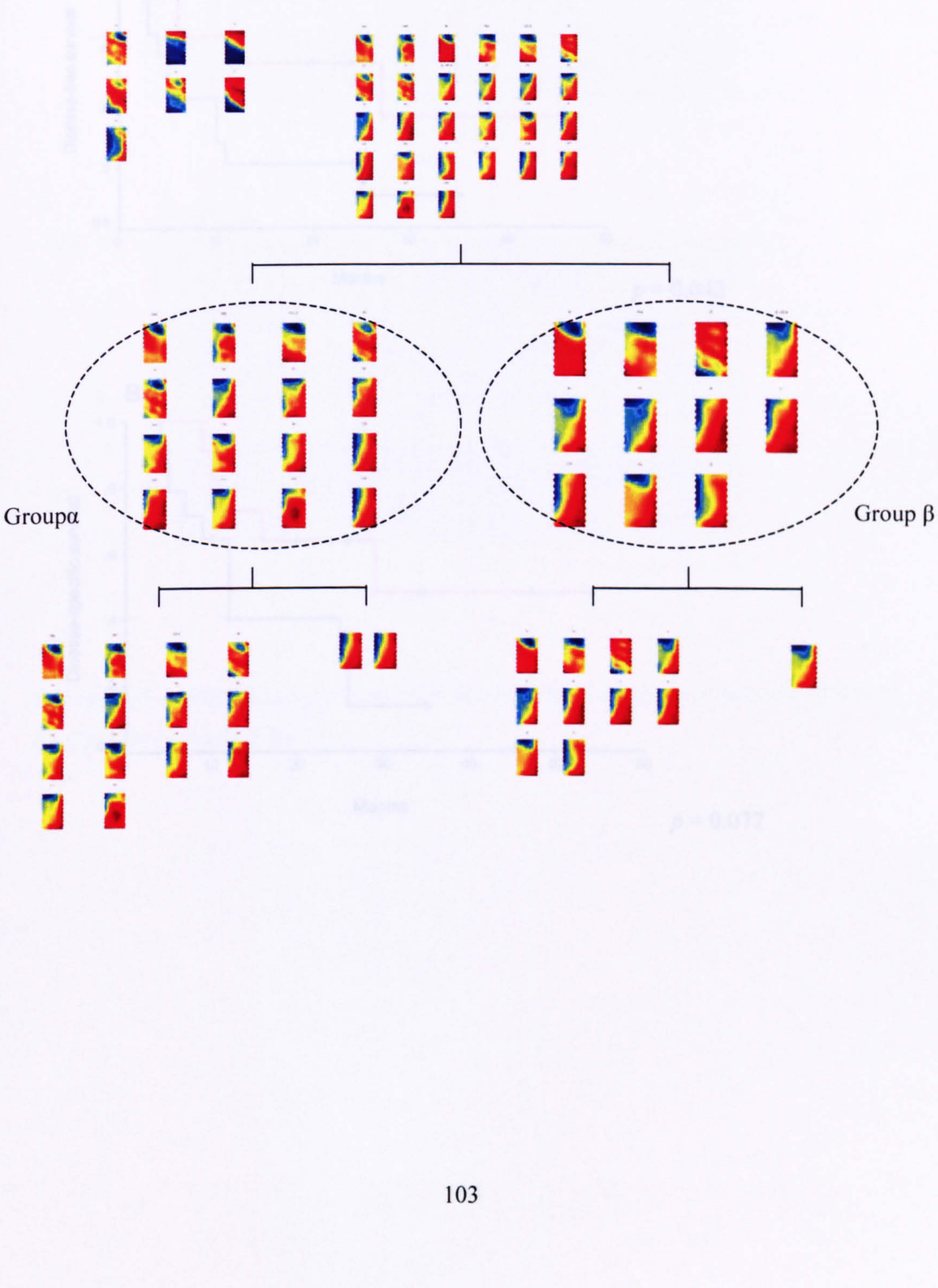
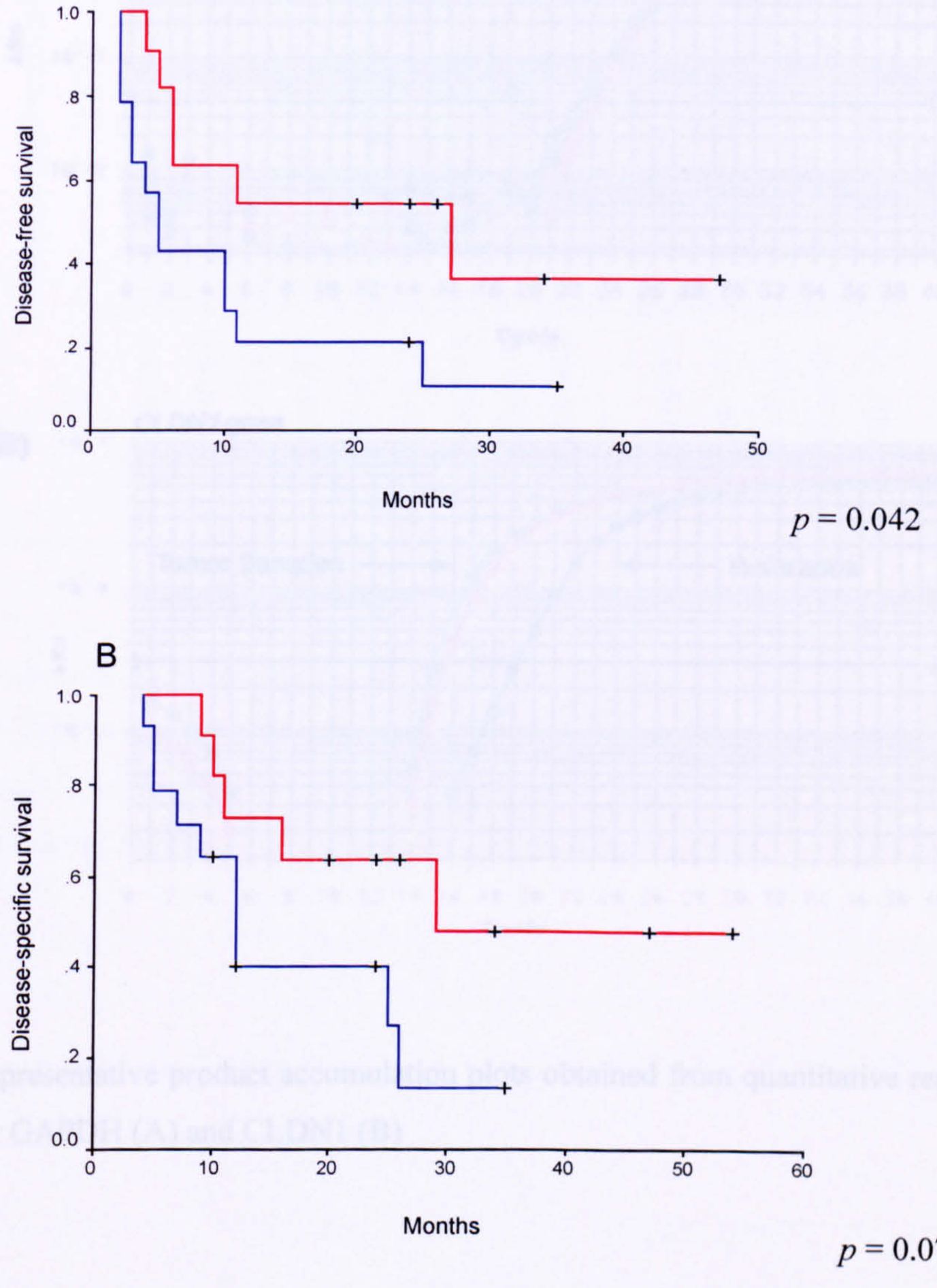


Figure 7

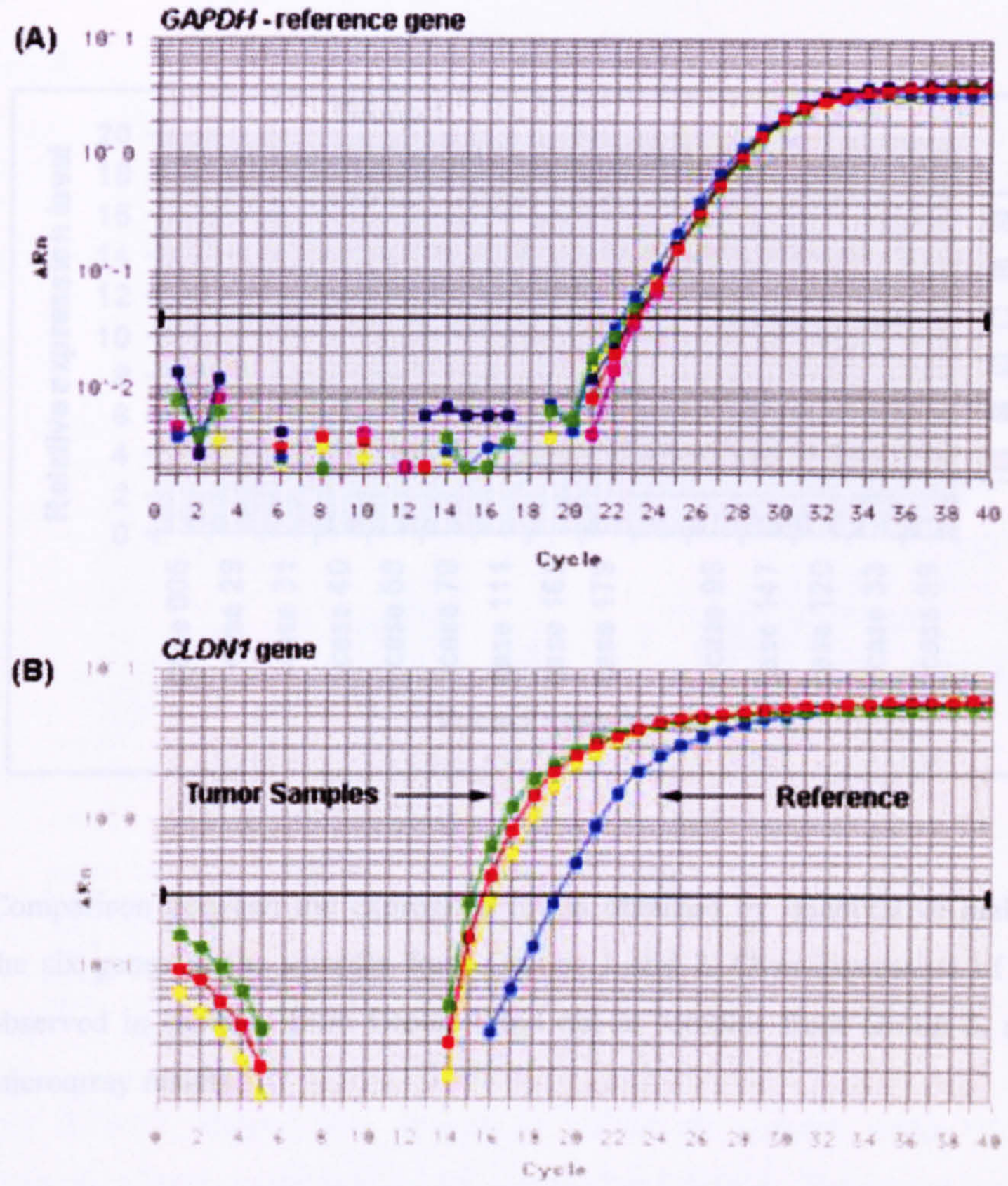
Disease-free (A) and disease-specific (B) Kaplin-Meier survival graphs for Group α (bottom line) and Group β (top line). + censored patients



$p = 0.042$

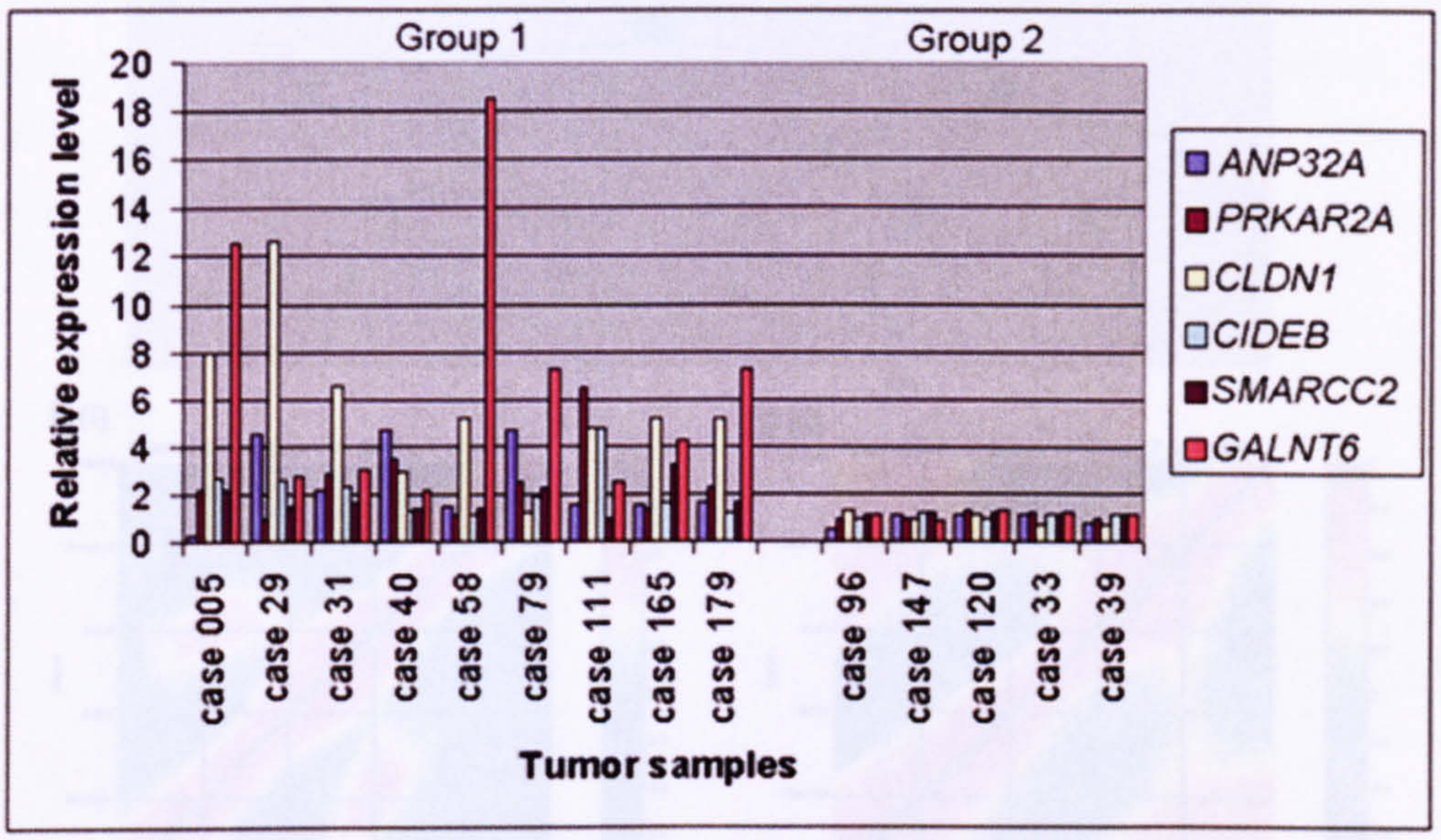
$p = 0.077$

Figure 8



Representative product accumulation plots obtained from quantitative real-time RT-PCR for GAPDH (A) and CLDN1 (B)

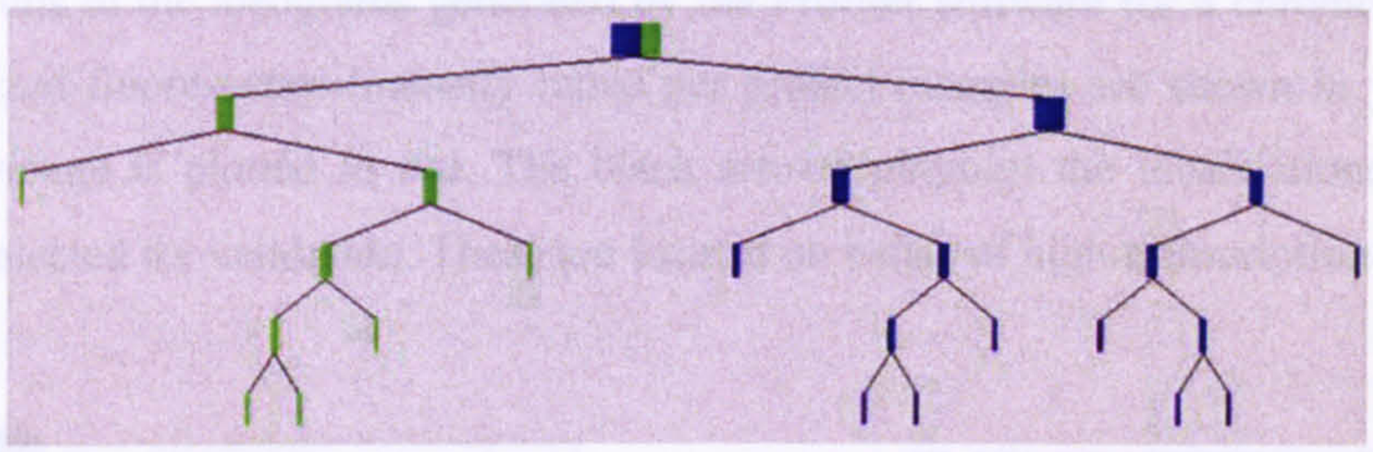
Figure 9



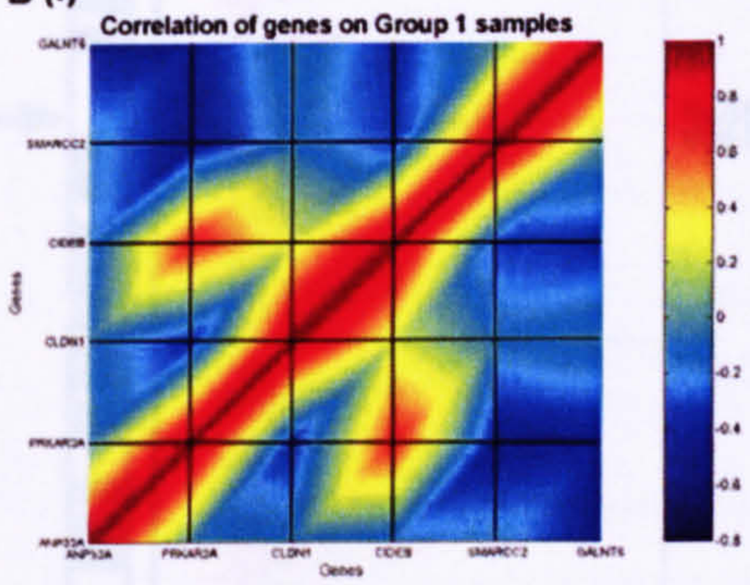
Comparison between the expression levels obtained by quantitative real-time PCR for the six genes in the samples from Groups 1 and 2. Over-expression of these genes is observed in samples from Group 1 but not in samples from Group 2, consistent with microarray results.

Figure 10

A

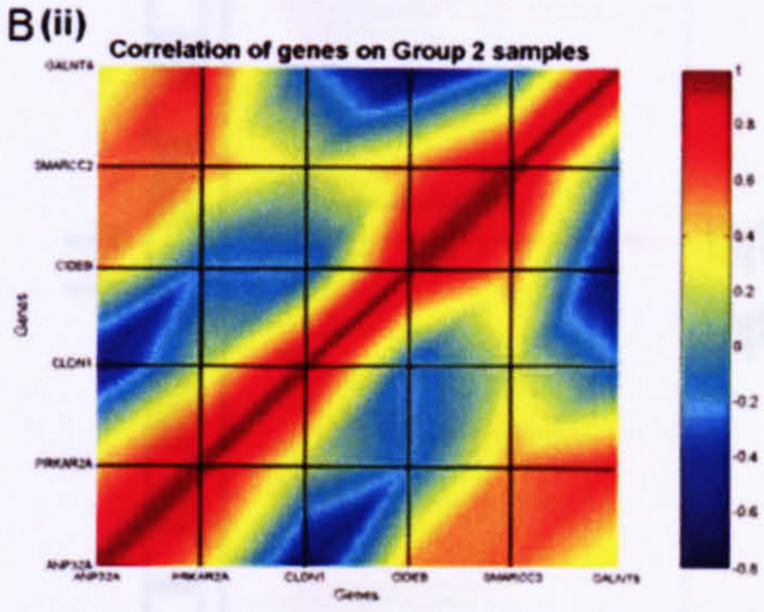


B (i)



GALNT6	-0.4647	-0.4367	-0.0598	-0.3765	0.0525	**1.0000**
SMARCC2	-0.1749	-0.4728	-0.1270	-0.3287	**1.0000**	0.0525
CIDEB	-0.1961	**0.7120**	0.2436	**1.0000**	-0.3287	-0.3765
CLDN1	-0.1192	-0.3747	**1.0000**	0.2436	-0.1270	-0.0598
PRKAR2A	-0.0605	**1.0000**	-0.3747	**0.7120**	-0.4728	-0.4367
ANP32A	**1.0000**	-0.0605	-0.1192	-0.1961	-0.1749	-0.4647
	ANP32A	***PRKAR2A***	***CLDN1***	***CIDEB***	***SMARCC2***	***GALNT6***

B (ii)

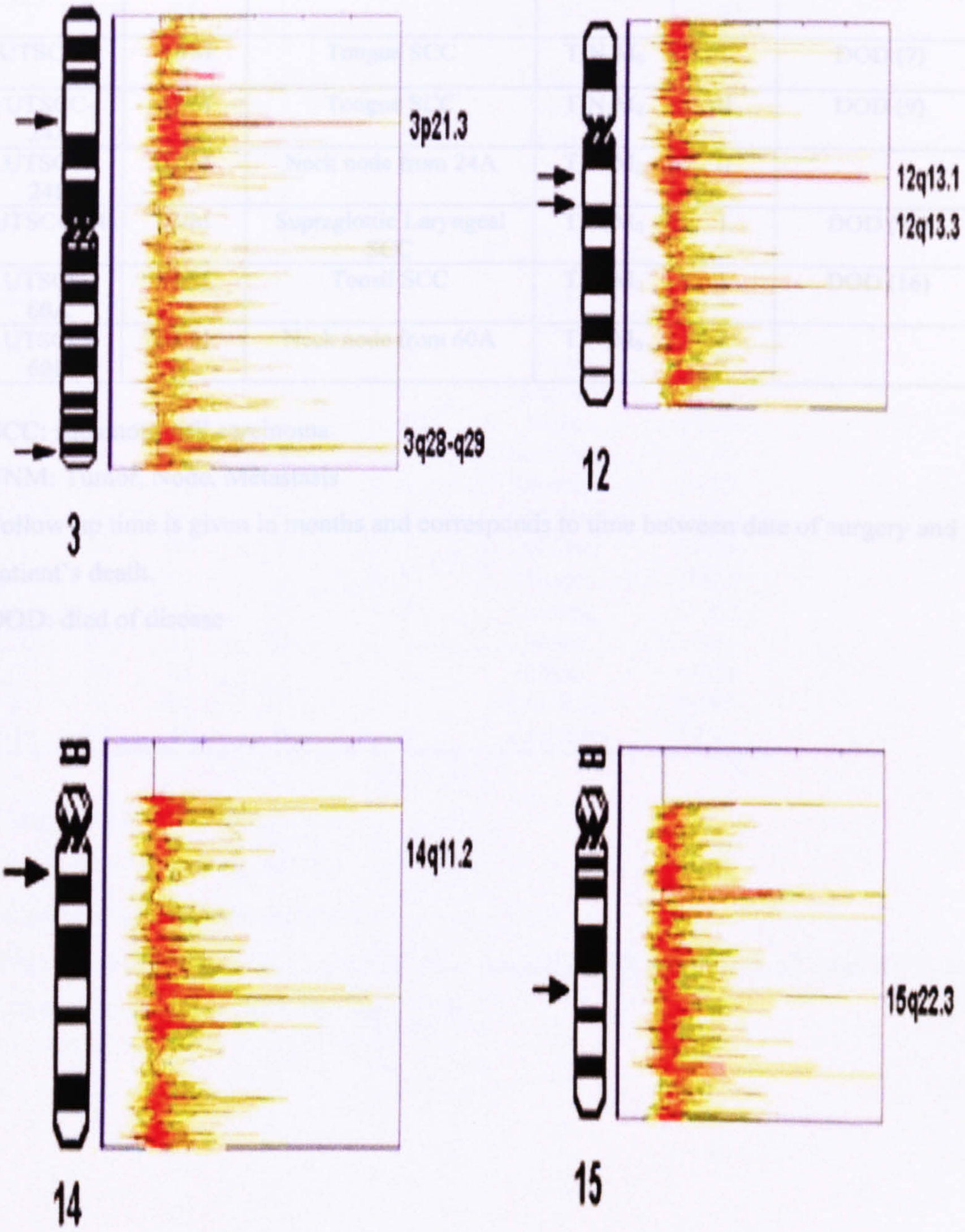


GALNT6	0.2176	**0.7097**	-0.1972	-0.6639	-0.1631	**1.0000**
SMARCC2	0.5702	**0.5298**	0.2287	0.5227	**1.0000**	-0.1631
CIDEB	0.5146	-0.2001	-0.2119	**1.0000**	0.5227	-0.6639
CLDN1	-0.6309	0.1692	**1.0000**	-0.2119	0.2287	-0.1972
PRKAR2A	0.4491	**1.0000**	0.1692	-0.2001	0.5298	**0.7097**
ANP32A	**1.0000**	0.4491	-0.6309	**0.5146**	0.5702	0.2176
	ANP32A	***PRKAR2A***	***CLDN1***	***CIDEB***	***SMARCC2***	***GALNT6***

(A) BTSVQ analysis was applied to cluster all samples using relative transcript expression levels as determined by quantitative RT-PCR. The binary tree clearly shows that the six genes are highly predictive, as all samples from Group 1 (blue) are split from Group 2 (green) samples at the first level of the tree. (B) Pseudo-color correlation matrix Group 1 samples and Group 2 samples. The *color map* corresponds to the scale of correlation coefficients: non-correlated data show with a coefficient of zero (*light blue*), negative correlation *dark blue*, and positive correlation ranging from *yellow* to *red*. The diagonal of the symmetric correlation matrix represents self-correlation and thus is equal to one (*dark red*). The numerical correlation co-efficients are shown below.

Figure 11

The results of the ideograms generated by the Profiler software for 4 chromosomes. The normalized fluorescence intensity ratios per project (sample) are shown in yellow, and their average is plotted in red. The black arrows pinpoint the localizations of the six genes selected for validation. These are located on bands of high transcriptional gain.



TABLES:

Table I.

Clinical data of patients from whom cell lines were derived.

Cell line	Age/Sex	Tumor site	TNM stage	Grade	Outcome (Follow-up time)
UTSCC-9	81/M	Tongue SCC	$T_2N_1M_0$	I	DOD (7)
UTSCC-24A	41/M	Tongue SCC	$T_2N_1M_0$	II	DOD (9)
UTSCC-24B	41/M	Neck node from 24A	$T_2N_1M_0$	II	
UTSCC-34	63/M	Supraglottic Laryngeal SCC	$T_4N_0M_0$	I	DOD (10)
UTSCC-60A	59/M	Tonsil SCC	$T_4N_1M_0$	I	DOD (16)
UTSCC-60B	59/M	Neck node from 60A	$T_4N_1M_0$	I	

SCC: squamous cell carcinoma

TNM: Tumor, Node, Metastasis

Follow-up time is given in months and corresponds to time between date of surgery and patient's death.

DOD: died of disease

Table II:

Patient demographics and tumor details

ID	Age	Sex	T Stage	N Stage	Recurrence	Outcome
79	44	M	1	1	Loc/Reg	NED
111	46	M	1	1	Loc/Reg	DOD
34	58	M	1	0	Distant	DOD
58	58	M	0	1	Loc/Reg	NED
63	59	M	1	1	Loc/Reg	DOD
39	60	F	1	0	Loc/Reg	AWD
165	62	F	1	1	Distant	DOD
29	64	M	1	1	None	NED
40	66	M	1	1	None	NED
162	74	F	1	0	None	DOC
96	75	M	1	1	Loc/Reg	DOD
45	37	F	1	0	None	NED
2	60	M	1	1	Distant	DOD
85	64	M	0	0	None	DOC
25	65	M	1	0	None	NED
193	75	M	0	0	None	DOC
19	48	M	1	0	Loc/Reg	DOD
131	56	M	1	0	None	NED
26	66	M	0	0	Loc/Reg	DOD
179	69	M	1	1	Distant	DOD
125	26	F	0	1	Loc/Reg	DOD
31	48	M	1	1	Loc/Reg	DOD
147	50	F	0	1	None	NED
86	53	M	1	0	Loc/reg	DOD
120	67	F	0	0	None	NED
33	70	M	1	0	None	NED
5	74	M	1	1	Loc/Reg	DOD
129	83	F	0	0	Loc/Reg	DOD

T stage, 0 = I-II, 1 = III-IV

N stage, 0 = NO 1 = N1, N2, N3

Loc/Reg = Locoregional recurrence

DOD, Died of disease; DOC, Died of other causes; AWD, Alive with disease; NED, No evidence of disease at last follow up

Table III

Description of patients in Group1 and 2 as defined by BTSVQ analysis of expression data from 20 OSCC samples, clinical and histopathologic data, local or distant recurrence and outcome.

Case number	Age/Sex	Tumor site	T category	N category	Recurrence *(months)	Outcome *(months)
Group 1						
34	58/M	FOM	3	0	Distant (11)	DOD (11)
40	66/M	FOM	3	1	No	NED (24)
58	58/M	FOM	2	1	Loc/Reg (10)	NED (12)
63	59/M	FOM	3	1	Loc/Reg (2)	DOD (5)
162	74/F	Tongue	4	0	No	DOC (26)
005	74/M	Tongue	4	1	Loc/Reg (6)	DOD (6)
79	44/M	FOM	4	1	Loc/Reg (4)	NED (54)
165	62/F	Tongue	4	1	Distant (25)	DOD (26)
179	69/M	Tongue	3	1	Distant (11)	DOD (12)
31	48/M	Tongue	4	1	Loc/Reg (27)	DOD (29)
111	46/M	FOM	4	1	Loc/Reg (7)	DOD (7)
29	64/M	Tongue	4	1	No	NED (24)
Group 2						
86	53/M	Tongue	3	0	Loc/Reg (5)	DOD (9)
96	75/M	FOM	4	1	Loc/Reg (3)	DOD (5)
120	67/F	Tongue	2	0	No	NED (34)
129	83/F	Tongue	2	0	Loc/Reg (2)	DOD (4)
33	70/M	FOM	4	0	No	NED (20)
39	60/F	FOM	4	0	Loc/Reg (10)	AWD (10)
147	50/F	Tongue	2	1	No	NED (35)
125	26/F	Tongue	2	1	Loc/Reg (2)	DOD (25)

FOM: Floor of Mouth

N category, 0 = N0, 1 = N1, 2a, 2b, 2c, 3

Loc/Reg = Locoregional recurrence

No= No recurrence

DOD, Died of disease; DOC, Died of other causes; AWD, Alive with disease; NED, No evidence of disease at last follow up.

*: time to recurrence and outcome in months.

Table IV.

Correlation of tumor characteristics, patient demographics and outcome based on BTSVQ defined groups from 20 OSCC samples.

	Group 1	Group 2	*p* value
Sample Size	12	8	
Male gender	10	3	0.035[a]
Median Age	60	63	0.62[b]
T stage			
I-II	1	4	0.035[a]
III-IV	11	4	
N stage			
Node –ve	2	5	0.035[a]
Node +ve	10	3	
Recurrence	9	5	0.55[a]
2-year disease-free survival	58%	62%	0.42[c]

Node –ve: absence of lymph node metastasis

Node +ve: presence of lymph node metastasis

[a] Fisher's exact test

[b] Mann-Whitney U test

[c] Log-Rank Test

Table V:

Comparison of patient demographics, survival and tumor characteristics between the two groups defined by BTVSQ from the analysis of 34 samples.

	Group α	Group β	P
Sample Size,*n*	16	11	
Male gender,*n*	10*	9	0.39[a]
Median Age	59*	63	0.42[b]
T stage,*n*			
I-II	6*	1	0.79[a]
III-IV	9*	10	
N stage,*n*			
Node –ve	3*	7	0.024[a]
Node +ve	12*	4	
Recurrence,*n*	12*	6	0.085
Median follow-up, months	12*	24	
2-year disease-free survival	21%*	55%	0.042[f]
2-year disease-specific survival	42%*	64%	0.077[f]

[a] Chi-square test

[b] Mann-Whitney U test

* Calculated from 15 patients

[f] Log-Rank Test for the comparison of Kaplan-Meier survival curves.

Table VI

Significant deregulated genes identified by BTSVQ analysis of 20 OSCC samples. Validated genes are in bold. * Function not described in the available sources of information on genes and proteins

Unigene ID	Gene symbol/name	Cytogenetic location	Function
Hs.188614	EST	Unknown	Unknown
Hs.142442	HP1-BP74	1p36.13	DNA binding activity, nucleosome assembly
Hs.13144	*ORMDL2/ORM*1-like2	12q13.13	*Known or inferred
Hs.271692	EST	Unknown	Unknown
Hs.285013	***ANP32A*/acid (leucine-rich) nuclear phosphoprotein 32 family, member A**	**15q22.33**	**Intracellular signaling cascade**
Hs.7327	***CLDN1*/claudin 1**	**3q28-q29**	**Cell-cell adhesion in epithelial endothelial cells**
Hs.42722	EST	Unknown	Unknown
Hs.365523	***PRKAR2A*/protein kinase, cAMP-dependent, regulatory, type II, alpha**	**3p21.3**	**c-AMP dependent protein kinase, intracellular signaling cascade**
Hs.343244	*AP1G2*/adaptor-related protein complex 1, gamma 2 subunit	14q11.2	Transport of ligand-receptor complexes from the plasma membrane or from the trans-Golgi network to lysosomes
Hs.187505	EST	Unknown	Unknown
Hs.75360	*CPE*/carboxipeptidase E	4q32.3	Cleaves C-terminal amino acid residues and is involved in neuropeptide processing
Hs.151678	***GALNT6*/UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 6**	**12q13.13**	**Catalyze acetylgalactosaminyltransferase reactions**
Hs.42743	Hypothetical	11q21	Unknown
Hs.193698	EST	Unknown	Unknown
Hs.12294	EST	Unknown	Unknown
Hs.236030	***SMARCC2*/SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily c, member 2**	**12q13.3**	**Transcription regulation and coactivation, chromatin remodeling**
Hs.12144	KIAA1033	12q24.11	Unknown
Hs.31532	Homo sapiens mRNA; cDNA DKFZp434F172 (from clone DKFZp434F172)	Unknown	Unknown
Hs.82001	*PKD2*/polycystic kidney disease 2 (autosomal dominant)	4q21-q23	Calcium-activated intracellular calcium release channel in vivo
Hs.156016	KIAA0140	Unknown	Unknown
Hs.49391	c21orf91/chromosome 21 open reading frame 91	21q21.1	Unknown
Hs.138485	EST	Unknown	Unknown
Hs.288835	***CIDEB*/cell-death inducing DFFA-like effector B**	**14q11.2**	**Cell death, induction of apoptosis by DNA damage**

Table VII

Selection of genes with a QE score <0.1 that discriminated Group α (Cluster 1) and β (Cluster 2) generated by BTSVQ analysis of 34 samples.

	GeneBank accession no.	Unigene ID	Gene annotation
Cluster 1	W90717	Hs.177386	EST
	R49227	Hs.13702	EST
	N34216	Hs.33519	EST
	R62994	Hs.78575	Prosaposin
	N48062	Hs.13809	FLJ10648
	N74025	Hs.251415	Human Type I iodothyronine-deiodinase
	N69121	In multiple clusters	
	AA131920	Hs.214368	EST
	R98073	Hs.172382	FLJ20001
	R06909	Hs.269029	EST
	R61019	Hs.1001855	EST
	H00517	In multiple clusters	
	W88532	Hs.254562	EST
	AA203290	Hs.25648	Tumour necrosis factor receptor superfamily, member 5
	W93973	Hs.77572	BCL2/adenovirus E1B 19kD-interacting protein 1
Cluster 2	H65659	Hs.100009	Human peroxisomal acyl-CoA oxidase
	H53489	Hs.1012	Complement component 4-binding protein α
	W37850	Hs.79101	Cyclin G1

Table VIII

Primer sequences used for validation by quantitative real-time RT-PCR.

Gene	Primers Forward (F) and Reverse (R)
CIBEB	F- 5'-GTCCTCTGATCCCCTCGTGA-3' / R- 5'-GGGATGTGAGGCGTTATGCT-3'
PRKAR2A	F- 5'-AAAGGATGGGCAGAGGTTCA-3' / R- 5'-GGGCCTTCAGAAGCAAAGTG-3'
ANP32A	F- 5'-GGGACATTCCCCATCTCTCA-3' / R- 5'-CCCACCACCATCTGTGAAGG-3'
SMARCC2	F- 5'-GTGGCTCCAGCCTCTGTAGT-3' / R- 5'-GTTCAGAAGGGCCCAAACTT-3'
CLDN1	F- 5'-GCCCCAGTGGAGGATTTACT-3' / R- 5'-GCAATGTGCTGCTCAGATTC-3'
GALNT6	F- 5'-CACCGATGGAAGAGACCATT-3' / R- 5'-TCCCTACTTTGGGAGCCTCT-3'
GAPDH	F- 5'-GGCGACGCAAAAGAAGATG-3' / R- 5'-CCGTTGACTCCGACCTTCAC-3'

Table IX

Relative expression levels as detected by quantitative real-time RT-PCR for the six genes in samples from Group 1 and 2.

Case number			Genes			
Group 1	SMARCC2	CIDEB	CLDN1	ANP32A	GALNT6	PRKAR2A
5	2.235	2.657	7.996	0.330	12.550	2.143
29	1.459	2.629	12.668	4.469	2.789	0.959
31	1.658	2.403	6.590	2.151	3.125	2.878
40	1.385	1.310	2.958	4.680	2.234	3.506
58	1.375	1.098	5.207	1.510	18.507	1.109
79	2.245	1.847	1.230	4.659	7.308	2.534
111	1.010	4.806	4.815	1.575	2.479	6.498
165	3.149	1.664	5.159	1.607	4.287	1.399
179	1.687	1.177	5.205	1.723	7.310	2.310
Mean ± *SD	1.800 ± 0.645	2.176 ± 1.155	5.758 ± 3.233	2.522 ± 1.633	6.732 ± 5.545	2.592 ± 1.686
Group 2						
33	1.052	1.085	0.712	1.046	1.240	1.161
39	1.046	1.046	0.740	0.840	1.084	0.949
96	1.051	1.026	1.280	0.450	1.060	1.043
120	1.215	1.043	1.057	1.052	1.250	1.337
147	1.206	1.222	0.949	1.049	0.900	1.035
Mean ± *SD	1.114 ± 0.088	1.084 ± 0.079	0.947 ± 0.235	0.887 ± 0.260	1.106 ± 0.144	1.105 ± 0.150

*SD: Standard Deviation

Table X.

Correlation between clustering of samples in Groups 1 and 2 and relative expression levels of genes, as detected by quantitative real-time RT-PCR.

Category			Genes			
	CLDN1	SMARCC2	GALNT6	ANP32A	CIDEB	PRKAR2A
Group1 vs. Group 2	p=0.007*	p=0.013	p=0.016	p=0.017	p=0.022	p=0.077

Adjusted p value using Bonferroni correction: $p \leq 0.0083$.